A Critical Review of Measures of Mentalization from Peter Fonagy's Conceptual Framework

by

Carla Rumeo

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirements for the degree of

Master of Arts

in

Psychology

Waterloo, Ontario, Canada, 2022

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Background: Mentalization is an expansive and multifaceted construct with important implications for psychological research as well as the etiology and treatment of psychological disorders. It is defined as the process through which an individual infers their own and others' mental states are intentional and lead to meaningful actions (Bateman & Fonagy, 2004). The theoretical framework in which mentalization is situated has evolved over the years and is widely accepted however, the empirical work surrounding the measurement and operationalization of mentalization is not as well developed and merits further investigation**. Methods:** The authors examined the psychometric soundness and construct validity of the gold standard, interview-based measure of mentalizing as well as five self-report scales. To assess the convergent validity of these measures, Canadian university students (N = 247) completed three self-report measures of mentalization as well as one task-based tool, the Movie for the Assessment of Social Cognition (MASC; Dziobek et al., 2006). To investigate whether self-report measures predict performance on the MASC, twenty linear regressions were estimated. Exploratory factor analysis was conducted to identify the common latent factors underlying all five self-report measures at the subscale level. **Results:** Certain self-report measures were strongly linked and common content included items about emotion recognition and regulation, understanding one's motivations for their behaviors and making accurate inferences about others' thoughts. Other measures were weakly correlated and dissimilar in item content. All self-report measures were weakly correlated with the MASC. Most of the regression models were non-significant. Of the four models that emerged as significant and had significant direct effects, a classical suppression effect was observed, which merits replication. Exploratory factor analysis revealed a one factor solution fit the subscale level data well. **Conclusions:** This study provides preliminary evidence

that there is some convergence among self-report measures. Researchers should reflect on their choice of instrument and should not use different tools interchangeably. The lack of convergence between task-based and self-report measures is disconcerting and warrants further research in this area. Last, whether the latent construct of mentalization is indeed unidimensional or has a more complex factor structure is yet to be determined.

Keywords: mentalization, theory of mind, empathy, self-report measures, task-based measures, construct validity, convergent validity

Acknowledgements

There were multiple parties who played prominent roles in the completion of this thesis, all of whom I am fortunate to have as part of my personal and professional support system. I am extremely grateful for the guidance provided by my research supervisor, Dr. Jonathan Oakman, whose breadth of statistical and theoretical knowledge and unwavering patience, flexibility and support enabled me to pivot this project not once, but twice. I would also like to thank my second readers, Dr. Uzma Rehman and Dr. Liz Nilsen for devoting their time and energy to provide insightful comments and feedback, which facilitated the process of finalizing this project and brought about important suggestions for future work.

This project could not have come to fruition without the unconditional love, support, and kindness provided by my parents, Ignazio and Lucia, who have encouraged me to pursue my aspirations since I was a child. Your continued guidance and support give me the inspiration and direction to continue to work towards my professional goals! To my brother, Alessandro, you never fail to put a smile on my face, even during the most stressful and difficult times. To my partner, Timour, you've been with me through it all these last 10 years and I am so appreciative of your continued support and encouragement throughout my journey. The goofiness and lightheartedness that we share adds a special levity to our life together amidst our rigorous professional lives. I am truly grateful to have you by my side, honey! Last, but certainly not least, I want to thank my friends, both from childhood and those that have recently come into my life. I want to specifically name my best friend Anna who has been immensely supportive of my academic and professional goals and who always encourages me to persist through hardships. To my cohort, it is surreal to think that we were only just meeting 2 years ago at this time given how close we've already become. I am so appreciative to be on this wild journey with all of you!

Table of Contents

List of Tables

Literature Review

An Overview of Mentalization

Mentalization is an expansive construct with important implications for psychological research as well as the etiology and treatment of mental health disorders. Peter Fonagy coined the term 'mentalization', defining it as "the capacity to conceive of conscious and unconscious mental states in oneself and others", incorporating tenets from psychoanalytic traditions with the construct of theory of mind (ToM; Allen, 2006; Choi-Kain & Gunderson, 2008). In his initial conceptualization, Fonagy proposed that impairments in mentalization (also described as defenses against thinking about mental states) may be a hallmark of individuals with Borderline Personality Disorder (Fonagy, 1991). Subsequently the scope of this construct has expanded and research has suggested that impairments in mentalizing are linked to a host of psychological disorders such as depression (Belvederi Murri et al., 2017; Bressi et al., 2016), anorexia nervosa (Ponti et al., 2019), and personality disorders such as antisocial and narcissistic personality disorder (Luyten et al., 2020). Many researchers have found evidence for the link between improvements in mentalizing throughout psychotherapy with positive changes in symptomology (Antonsen et al., 2016; Ekeblad et al., 2016; Euler et al., 2022; McGown et al., 2021), suggesting broad applicability of this construct. Moreover, Bateman et al. (2018) suggest that while improved mentalizing is a shared outcome across psychotherapeutic interventions, improvements in this skill are a mechanism of change in treatments specifically aimed at treating personality disorders. In addition to this area of interest, three other lines of mentalization focused research have burgeoned: 1) the etiological path of one's capacity to mentalize, 2) theoretical examinations of the conceptual boundaries of mentalization, and the 3) development of various measurements of this nuanced construct.

Despite the wide-ranging applicability of this construct and intriguing theoretical developments, the empirical work surrounding the measurement and operationalization of mentalization is not well-developed and merits further investigation. Additionally, the term "mentalization" lacks conceptual clarity and is used to describe a broad array of constructs (e.g., empathy, theory of mind, etc.) and types of social cognition tasks (e.g., emotion recognition and face perception). Whether or not these constructs are synonymous with mentalization or mere components of mentalization is yet to be determined. The current study will provide a critical review of five self-report measures and examine how well each measure aligns with Fonagy's conceptual framework.

Mentalization as an Umbrella Term

Mentalization is a multifaceted and broad concept. As such, many authors have pointed out that there is considerable conceptual overlap with other similar constructs such as theory of mind, psychological mindedness, mindfulness, emotional intelligence, and empathy (see Decety & Jackson, 2004 for a comprehensive review of empathy). Researchers have also highlighted distinct properties of mentalization. For instance, while mindfulness and mentalization both involve observing one's inner states, mindfulness is predicated on focusing on the "here and now" or the present moment whereas an individual can mentalize inner states from the past, present, or future. Moreover, one of the goals of mindfulness is to accept one's inner states as they are (thoughts, emotions, ideas, etc.) whereas mentalization involves appraising these states to derive meaning from others' actions (Choi-Kain & Gunderson, 2008).

Empathy has been defined in numerous ways, but most definitions share three common elements: 1) sharing another person's affective response 2) perspective taking, and 3) a stable capacity to differentiate between one's own and others' internal states (Choi-Kain & Gunderson,

2008). Both mentalizing and empathy involve sharing the affective experience of another person, yet empathic processes are primarily affectively based and oriented to others' emotional states (Choi-Kain & Gunderson, 2008; Katznelson, 2014; Luyten et al., 2020). Some researchers have maintained that both theory of mind (ToM) and mentalization are multidimensional constructs composed of affective and cognitive components and that adaptive mentalization and adequate ToM capacities require integration of these processes (Luyten et al., 2020; Sebastian et al., 2012; Shamay-Tsoory et al., 2010). It has been suggested that both mentalization and ToM incorporate one's knowledge of others' cognitive states through processes such as perspective taking and belief-desire reasoning (Choi-Kain & Gunderson, 2008; Katznelson, 2014; Luyten et al., 2020). However, affective components of mentalization and ToM may differentiate these broad constructs. Affective ToM involves identifying, understanding, and empathizing with others' emotional states (Sebastian et al., 2012; Shamay-Tsoory et al., 2010) and while the affective component of mentalization includes these processes, affective mirroring (e.g., feeling another person's emotions) is also a vital component of mentalizing others' emotional states. In contrast to the multidimensional framework for conceptualizing ToM, other authors assert that theory of mind skills primarily involve identifying and understanding cognitive states, with little emphasis on inferring others' affective states (Apperly, 2012; Choi-Kain & Gunderson, 2008). Last, it is possible that mentalizing represents the developmental progression of theory of mind skills in adulthood.

Another construct that has been theorized to overlap with mentalization is *affect consciousness* which "refers to the mutual relationship between activation of basic affective experiences and the individual's capacity to consciously perceive, tolerate, reflect upon and express these experiences" (Solbakken et al., 2011, p. 5) and "is the process thought to underlie

integration of affect in cognition, motivation, and behavior" (Falkenström et al., 2014, p. 2). Affect consciousness involves the same processes for understanding affect as described in relation to another similar construct, *mentalized affectivity* or one's capacity for emotion regulation while concurrently experiencing that affect (Fonagy et al., 2002; Jurist, 2005). These processes include identifying or labelling affect(s) one is currently experiencing, processing or modifying one's affect (e.g., modifying the intensity of the affect), and expressing one's affect both in a typical outward fashion as well as inward expressions which involves reflecting on one's affect (Solbakken et al., 2011). In terms of conceptual similarities, the concepts of mentalization and affect consciousness both place emphasis on the awareness, reflection, and expression of affect related internal states (Choi-Kain & Gunderson, 2008) as well as the implicit influence of affective states on one's perceptions and behaviors (Solbakken et al., 2011). While both mentalization and affect consciousness refer to reflective capacities pertaining to one's own as well as others' mental states, the conceptualization of mentalization provides more of an explicit focus on interpreting others' mental states whereas affect consciousness focuses more on understanding one's own affect and using this knowledge to make inferences about others to better understand them (Solbakken et al., 2011). Overall, it is evident that mentalization is conceptually similar to other constructs but also has distinct properties.

Development of Mentalization

In further theoretical developments, Fonagy and colleagues have proposed an etiological path for the development of this capacity. They posit that all individuals have an evolutionarily based and prewired capacity for mentalization which is then modulated by one's social and learning environment, especially early parent-child relationships (Fonagy & Luyten, 2016). Mentalization is thought to develop in the context of secure attachment relationships and through

parental mentalizing, commonly referred to as parental reflective functioning (PRF) (Luyten et al., 2020). PRF is defined as mentalizing in relation to the internal states of one's child. Caregivers with high PRF are thought to respond to children's affect with a "marked and contingent" response that mirrors the child's emotion but also slightly modifies it so that it may be differentiated (Choi-Kain & Gunderson, 2008, p. 1129). This process helps the child to translate the somatic, sensory experience into an internal, mental experience and thus fosters the child's capacity for mentalization. Luyten and colleagues (2020) have expanded their model of the development of mentalization and further assert that PRF is influenced by one's family and neighborhood and the overall sociocultural context. PRF in turn affects children's degree of epistemic trust (one's capacity to appraise knowledge learned from others as personally salient and applicable to other contexts), which later impacts social learning, secure attachment, and predicts high levels of mentalizing.

In addition to the theories surrounding the development of mentalization, Fonagy and Target (1996) have also put forward a theory of psychic reality which identifies three modes of primitive mentalization that emerge early in childhood and are later abandoned and replaced by a more sophisticated capacity for mentalizing (Gagliardini et al., 2018). These modes may also re-emerge in contexts where one's ability to mentalize is hampered, such as in circumstances which elicit high physiological and emotional arousal states (Fonagy & Luyten, 2016). This theory also incorporates qualitative descriptors of problematic thought patterns that individuals with BPD tend to experience (Fonagy & Allison, 2012; Luyten et al., 2020). In the *psychic equivalence mode*, an individual equates their inner experiences with external reality; one's thoughts and feelings are experienced as real events instead of mental representations and thus emotions are experienced to an intense degree (Hausberg et al., 2012; Katznelson, 2014; Luyten et al., 2020).

Internal events such as thoughts are held with extreme conviction and individuals often fail to consider alternative perspectives. In *pretend mode,* internal events (e.g., thoughts and feelings) are severely disconnected from reality, resulting in overwhelming cognitive and emotional experiences that are disjointed from external events and may lead to dissociation (Fonagy & Allison, 2012; Luyten et al., 2020). This mode is also termed hypermentalizing or pseudomentalizing which is characterized by assessments of mental states that are driven by implicit or automatic mentalizing and as a result inferences about others' internal states are unfounded (e.g., deficits in belief-desire reasoning) and individuals tend to conflate their own and others' mental states (Fonagy & Luyten, 2016; Fonagy et al., 2015). Last, in the *teleological mode,* individuals tend to ignore mental states unless they are tied to observable and tangible goals. Moreover, only actions with a physical impact, such as self-harm behaviors, are considered indicators of one's true intentions or mental states. Individuals also tend to be certain of their inferences about others' mental states as they base these assumptions on concrete, physical evidence (Fonagy & Target, 1997; Katznelson, 2014). The teleological mode reflects a clear over-emphasis on externally based mentalizing and deficits in explicit/controlled or reflective and effortful mentalizing (Fonagy & Luyten, 2016).

Further Developing the Construct of Mentalization

In parallel to other researchers' interest in mentalization, Fonagy and his colleagues have continued to refine their conceptualization and have developed two tools for its measurement, the interview-based Reflective Functioning Scale (RFS; Fonagy et al., 1998), and the Reflective Functioning Questionnaire (RFQ; Fonagy et al., 2016), a self-report tool. Bateman and Fonagy (2004a) expanded their description and defined mentalization as "the mental process by which an individual implicitly and explicitly interprets the actions of himself or herself and others as

meaningful on the basis of intentional mental states such as personal desires, needs, feelings, beliefs, and reasons". This expanded definition in combination with behavioral studies highlight four important dimensions of mentalization: implicit and explicit types of processing (also referred to as automatic and controlled processing), mentalizing in reference to self and other, mentalizing about internal or external features of individuals, and mentalization emphasizing cognitive and affective content and processes (Choi-Kain & Gunderson, 2008; Luyten et al., 2020). Implicit/automatic processing involves fast, unconscious and reflexive assessments of one's own or others' mental states whereas explicit/controlled mentalizing involves deliberate, conscious, and effortful reflection on mental states. Mentalization can occur in the context of reflecting on *one's own* as well as *others'* internal states. Mentalizing also entails cognitive operations as well as affective processes and adaptive mentalizing is conceptualized as a balanced integration of both processes. Cognitive processes in mentalizing involve perspective taking, belief-desire reasoning (capability to predict an individual's actions based on interpreting their beliefs and desires), and an appreciation that one's thoughts are only one interpretation of reality amidst many other possibilities. The affective domain of mentalizing involves processing, recognizing, and expressing emotional experiences as well as the visceral sensation of another's emotions (e.g., affective mirroring). Last, individuals may aim to focus on others' *internal* processes (e.g., by pondering their thoughts and emotions based on past behaviors and current factors) or they may focus on external cues such as others' facial expression, prosody of voice, and body posture for the purposes of inferring others' mental states (Luyten et al., 2020).

Psychopathology and Mentalization from a Multidimensional Model

Maladaptive mentalizing emerges when there is an imbalance in the four dimensions highlighted above; different psychological disorders are associated with various constellations of

imbalances in these dimensions (Luyten et al., 2020). Specifically, imbalances refer to excessive functioning of one polarity (e.g., affective processing) and a corresponding dysfunction in the opposite polarity (e.g., cognitive processing). While the notion that various imbalances are indicative of different disorders is a compelling theoretical stipulation, Gagliardini et al. (2018) have pointed out that there are conceptual issues concerning the idea that excessive reliance on one polarity is accompanied by deficits in the opposite polarity, particularly for the internal/external and controlled/automatic dimensions. For example, excessive reliance on internally based mentalizing does not necessarily imply a deficit in external mentalizing.

Despite these conceptual issues, Fonagy and colleagues (2015) have proposed a specific mentalization profile characterizing BPD. They assert that individuals with BPD are oriented towards automatic, externally, affectively, and self–related mentalization which has the following repercussions: impaired social cognition (e.g., being distrusting of others), implausible thoughts about others' feelings, beliefs, and intentions, and a lack of demarcation between one's own and others' mental states. Furthermore, Luyten et al. (2020) assert that this mentalizing profile can explain the empathy paradox (Dinsdale & Crespi, 2013) seen in individuals with BPD. They posit that quick, automatic and externally based mentalizing can sometimes result in accurate inferences about individuals' thoughts/beliefs, leading to superior performance in inferring others' intentions but may also lead to over-interpreting others' mental states and holding inaccurate conclusions with extreme conviction (e.g., individuals with BPD have been shown to negatively interpret neutral faces).

Aside from the predominantly theoretical work surrounding BPD, very few studies have empirically investigated whether different types of imbalances on these dimensions are indeed indicative of different disorders. Gagliardini et al. (2018) found that various types of opposing

imbalances (e.g., both self and other) were significant predictors of many personality disorders. These findings seem to contradict the notion that there are stable mentalizing profiles of imbalances that characterize different disorders. However, some researchers propose that mentalizing has both trait and state features. State-like features suggest that one's capacity to mentalize at a given moment is contingent on one's environment, arousal/stress level, and relationships with individuals that are present (Fonagy et al., 2015; Luyten et al., 2020). As such, specific types of mentalization imbalances may correspond to particular personality disorders, however these deficits may also fluctuate based on contextual factors. To the authors' knowledge, there is little to no research exploring this context-based approach to mentalizing imbalances and Luyten et al. (2020) suggest that further research using "dynamic modeling approaches is needed to further study the factors involved in explaining stability and changes in mentalizing across different contexts and time spans" (p. 305). In sum, while some stable mentalization profiles have been put forward (e.g., BPD), the theoretical tenet that various imbalances across the four dimensions are indicative of different psychological disorders is a compelling position but has yet to be thoroughly investigated. Furthermore, whether there are indeed contextual caveats to these mentalization profiles remains uncertain.

Psychopathology and Mentalization from a Dual Deficit Model

Complementing the theoretical propositions of the multidimensional model of mentalizing, Fonagy and other collaborators have maintained that there are two types of mentalizing impairments, hypomentalizing and hypermentalizing, both of which influence vulnerability to and/or maintenance of psychological disorders (Fonagy & Luyten, 2016; Fonagy et al., 2016; Sharp et al., 2011). *Hypomentalizing* refers to a severe lack of mentalization skills (e.g., unable to develop complex models which explain behavior in terms of mental states). In

contrast, *hypermentalizing* (or pseudomentalizing) refers to the generation of inaccurate, overly complex mental models to explain behavior. Specifically, these models are inferred through overinterpreting one's own or others' mental states without considering their social context and as such, these inferences are often unfounded (Fonagy et al., 2015; Fonagy et al., 2016). There is increasing empirical evidence supporting the notion that both hypomentalizing and hypermentalizing are related to psychopathology, with a larger evidence base supporting the role of hypomentalization. Various studies have found that individuals with psychological disorders (e.g., BPD, depressive disorders, eating disorders, and psychosis) had lower scores on mentalizing measures such as the Reflective Functioning Scale (RFS; Fonagy et al., 1998) in comparison to non-clinical populations (Ekeblad et al., 2016; Fischer-Kern et al., 2010; Gullestad et al., 2013; Levy et al., 2006; MacBeth et al., 2011; Ward et al., 2001). In contrast, some studies have found similar RF scores between clinical and non-clinical groups (Bressi et al., 2016; Karlsson & Kermott, 2006; Pedersen et al., 2012; Taubner et al., 2011; Staun et al., 2010). A few of these studies found no significant differences in mentalizing capacity between clinically depressed individuals and healthy controls (Staun et al., 2010; Taubner et al., 2011) or between outpatients with bulimia nervosa and a non-clinical group (Pedersen et al., 2012). These findings may be explained by some clinical samples having less severe pathology and a lower degree of impairment. For instance, studies which found no differences in mentalizing between clinical and non-clinical groups included outpatients who sought out psychotherapy treatment. In contrast, studies that have found that depressed individuals had a notably lower capacity for mentalizing involved participants who were in psychiatric inpatient units and had participated in previous treatments that were unsuccessful. Taken together, deficits in mentalization (e.g., hypomentalization) can be considered a general factor which may confer vulnerability to

psychopathology with variables moderating this association (e.g., severity and chronicity of symptoms, degree of functional impairment, previous experience with treatment, etc.).

In support of the latter tenet proposing that hypermentalizing is associated with the development and/or maintenance of psychopathology, a few studies have found that hypermentalizing is associated with symptomology of BPD as well as depression and anxiety disorders. Specifically, hypermentalizing errors on the Movie for the Assessment of Social Cognition (MASC; Dziobek et al., 2006), a task-based measure of mentalizing, were strongly associated with borderline traits in adolescents and explained further variance in predicting BPD when added to a logistic regression with internalizing, externalizing problems, and gender as predictors (Sharp et al., 2011). Hypermentalizing (as well as hypomentalizing) have also been found to mediate the relationship between experiences of emotional abuse in childhood and depressive symptoms in adulthood (Li et al., 2020). Ballespí et al. (2019) found that individuals with social anxiety disorder demonstrated hypermentalizing errors when asked to make inferences about neutral social signs demonstrated by an interviewer in a previous encounter. However, in contrast to Sharp et al.'s (2011) findings, social anxiety was not a predictor of errors on the MASC (Dziobek et al., 2006). Last, hypermentalizing was found to mediate the relation between trauma symptoms and self-reported aggression (Abate et al., 2017) in an adolescent inpatient population. In sum, there is some evidence supporting the role of both hypomentalization and hypermentalization as transdiagnostic vulnerability factors. It is worth highlighting that while various self-report and task-based measures assess hypomentalization, only two measures to date capture hypermentalizing: the RFQ (Fonagy et al., 2016) and the MASC (Dziobek et al., 2006). As such, it is possible that fewer research studies have investigated the role of hypermentalizing in psychopathology (in comparison to

hypomentalizing) and this may explain why less support has been garnered for hypermentalizing deficits.

Development of the Reflective Functioning Scale

Peter Fonagy's initial definition of mentalization, a capacity for thinking about conscious and unconscious mental states, provided the basis from which this concept continues to evolve today. Fonagy and colleagues (1998) also developed the first widely used measure of mentalization, the Reflective Functioning Scale (RFS). The term reflective function (RF) generally refers to the psychological processes that are implicated in mentalization capacities (Katznelson, 2014; Fonagy et al., 1998) although the terms RF and mentalizing seem to be used interchangeably throughout this research area. Fonagy and colleagues (1998) developed the RFS for the London Parent-Child Project, defining RF as a cognitive function which interprets behaviors of one's self and others in terms of internal mentals states such as beliefs, desires, etc. (Fonagy et al., 1998) Importantly, Fonagy et al. (1998) stated that the "reflective-functioning scale could be conceived of as providing operationalised definitions of individual differences in adults' metacognitive capacities" (Fonagy et al., 1998, p. 6). They were referring specifically to Mary Main's definitions of various metacognitive constructs, namely metacognitive knowledge or awareness that mental states are distinct from objective reality and thus, subject to fallibility. *Metacognitive knowledge* also involves acknowledging the dynamic and changing nature of mental states, that others may have different beliefs and desires than oneself, and that individuals' experiences are influenced by these internal states. Main described another related concept, metacognitive monitoring which "entails observation of, and curiosity about, the habits of mind, that shape our experience" (Wallin, 2007, p. 40-41).

The RFS is a coding system that can be applied to various clinical interviews but has mainly been used to code transcripts of the Adult Attachment Interview (AAI; George et al., 1984/1985/ 1996). The AAI requires 45-90 minutes to administer (Rutimann & Meehan, 2012) and includes 23 questions in which individuals are asked to elaborate on: their relationship with both of their parents when they were children, and specifically, their parents' responses when they were sick or in emotional distress, and their current relationship with their parents and their spouse/partner. Last, they are asked to consider how their experiences during childhood may impact their current (or imagined) parental behaviors in relation to their own children (Taubner et al., 2013a).

Psychometric Properties of the RFS Applied to the AAI

The RFS has been widely adopted in research on the etiology of psychopathology and the dynamic nature of caregiver-infant relationships and has demonstrated good psychometric properties across these studies. This scale was validated on a sample of 100 middle class mothers and fathers who completed the AAI prior to their first child's birth. Correlation coefficients between raters for mothers' RF was between .59 – .84 and coefficients between raters for fathers' RF was .79 – .89 (Fonagy et al., 1998). They also examined the associations between parents' RF and various demographic factors (e.g., level of education, gender, IQ, social class, socio-economic group, age, and years of cohabitation) and only found a modest correlation between fathers' RF and level of education (r = .35), suggesting that RF is minimally influenced by these factors. Taubner and colleagues (2013a) partially corroborated this work as they found that age and gender were not significant predictors of participants' RF scores. Moreover, recent investigations using the RFS have also yielded good levels of inter-rater reliability with Kappa scores ranging from .79 (Fischer-Kern et al., 2010) to .83 (Arnott & Meins, 2007) for overall RF

scores. Intraclass correlations (ICC) have been found to range from .71 – .90 (Berthelot et al., 2015; Diamond et al., 2014; Ensink et al., 2014; Falkenström et al. 2014; Gullestad et al., 2013; Levy et al., 2006; Pedersen et al., 2012; Taubner et al., 2013a). Taubner et al. (2013a) found that individual differences in RF scores were relatively stable over time, at both the item level (r = .34 to .47) and overall RF score (r = .64). The initial validation study (Fonagy et al., 1998) also illustrated good discriminant validity as RF was not significantly correlated with parental attitudes towards children (Mother-Father-Peer Scales; Epstein, 1983), personality traits such as extraversion and neuroticism (Eysenck Personality Questionnaire; Eysenck & Eysenck, 1975), symptoms of various psychological disorders (Langner's 22-item Mental Health Index; Langner, 1962), and self-esteem (Sources of Self Esteem Inventory; O'Brien, 1981). Falkenström et al. (2014) also found a significant, yet small correlation between global measures of mindfulness (Five Facet Mindfulness Questionnaire; Baer et al., 2006) and ratings on the RFS (r = .31), which supports the idea that these are related yet distinct constructs as noted by Choi-Kain and Gunderson (2008) and Woynowskie (2015).

While the RFS has shown promising psychometric qualities, Taubner et al. (2013a) performed a more in-depth analysis of the internal structure of the RFS coded from the AAI and examined the incremental validity of the demand question ratings in predicting the overall RF score. Taubner et al. (2013a) included two additional demand questions about individuals' current relationships with their parents and romantic partner. Confirmatory Factor Analysis (CFA) supported both a one-factor model and a two-factor model that distinguished between RF pertaining to past versus current relationships. However, since these factors were found to be highly correlated, the authors asserted that the unidimensional model may be superior given that it is more parsimonious. The authors also conducted a multiple regression analysis where each of

the demand question ratings were found to be significant predictors of the overall RF score; this model accounted for ~82% of the variance in the overall RF score. The authors investigated the incremental validity of the demand questions through partial correlations between each question and the global RF score while controlling for the other seven demand question ratings. They found significant yet low partial correlations for most of the demand questions ($.08 < r < .19$).

While Taubner et al. (2013a) concluded that a unidimensional model may be most suitable for the RFS, several other authors have taken issue with a 1-factor model and highlighted the utility of a multidimensional approach for assessing mentalization. Choi-Kain and Gunderson (2008) note that the construct of mentalization is multidimensional (i.e., interviews also take into account consistency, originality, and elaboration of speech) and that interviews can be given the same global score on the RFS yet discuss mental states in very different ways. Moreover, Luyten et al. (2020) describe mentalization as a multifaceted construct composed of four dimensions, but the RFS does not capture these dimensions.

Gullestad and Wilberg (2011) aimed to showcase the value of using a multidimensional approach to code RF and performed a qualitative analysis on two AAI's (at the pre-treatment and 36-month time period) for a single client during psychotherapy. The coders initially scored the interviews according to the RFS manual and later reread the interviews while attending to various dimensions of the client's RF (e.g., mentalizing about self vs. others and cognitively or affectively based mentalizing). They demonstrated that while the client's overall RF improved over the course of treatment, there were many more nuanced changes that were observed only in the latter approach (e.g., the client improved more on the cognitive than the affective component). These findings may suggest that the latter coding system may be better suited to investigations that examine mentalization as a potential mechanism of change in treatment.

Additionally, Taubner et al. (2011) found that a group of chronically depressed patients had similar global RF scores to healthy controls, yet individuals with chronic depression had lower RF scores on questions pertaining to experiences of loss. Taken together, these studies suggest that a global rating of RF may obscure important relationships between these lower order dimensions of mentalization.

Despite critiques of the unidimensional structure, construct validity of the RFS using this 1- factor structure has been showcased across multiple studies. In line with assertions that deficits in mentalization play a role in various psychopathologies, specifically BPD, Fonagy et al. (1996) found that RF discriminated between nonpsychotic psychiatric inpatients and matched controls with inpatients scoring much lower on the RFS. Scores on the RFS have also been found to be significantly correlated with number of axis I (r = -.29) and axis II (r = -.46) disorders (Bouchard et al., 2008). Additionally, individuals diagnosed with BPD in Fonagy and colleagues' (1996) study had significantly lower ratings of RF than patients diagnosed with anxiety, depression, substance use as well as antisocial and paranoid personality disorders. Other studies have also documented very low scores on the RFS (around 2-3) for borderline individuals (Fischer-Kern et al., 2010; Gullestad et al., 2013; Levy et al., 2006) as well as inpatients with major depressive disorder (Fischer-Kern et al., 2013), inpatients with anorexia nervosa (Ward et al., 2001), and patients experiencing their first psychotic episode (MacBeth et al., 2011). Taken together, these findings demonstrate construct validity of the RFS as they lend support for the notion that deficits in mentalization are a transdiagnostic factor influencing psychopathology (Luyten et al., 2020).

Moreover, predictive validity has been demonstrated through research that has suggested a link between high parental RF and their infants' secure attachment classifications (Fonagy et

al., 1991; Fonagy et al., 1994; Fonagy et al., 1998; Grienenberger et al., 2005; Slade et al., 2005; Stacks et al., 2014). Additionally, Slade et al. (2005) found that maternal RF largely accounted for the association between adult and infant attachment security, suggesting that parental RF may act as a mediator for the intergenerational transmission of attachment from parent to infant. Lending further credence to the construct validity of the RFS, Rosso et al. (2015) found that parental RF was positively correlated with their preadolescent children's mentalization capacities. These results align well with Fonagy and colleagues' theoretical work on the developmental path of mentalizing in children.

Alternatives Interviews for Measuring RF

In addition to the AAI, there are various interviews on which the RFS may be applied to index RF in different contexts, including the Parent Development Interview (PDI; Slade et al., 2004), Brief Reflective Function Interview (BRFI; Rudden et al., 2006), and Working Model of the Child Interview (WMCI; Zeanah & Benoit, 1995), to name a few. The PDI assesses a parent's RF in reference to their children and their representation of themselves as a parent. In the BRFI, interviewees are asked to elaborate on their childhood relationship with one parent/caregiver and to describe a current relationship with a nonparental attachment figure. In comparison to the AAI, the BRFI requires less time for administration (15 – 30 minutes), transcription and coding. The psychometric properties of the PDI and BRFI (e.g., inter-rater reliability and internal consistency) have been found to be adequate (Ensink et el. 2016; Rutimann and Meehan, 2012; Slade et al., 2005; Sleed et al., 2020; Stacks et al., 2014). Other parent adaptations of the RFS include the Caregiver Reflective Functioning Scale (CRFS; Gilbert, 2008) and the Maternal Reflective Functioning Scale (MRFS; Slade & Patterson, 2005) which is based on the Pregnancy Interview (Slade, 2003). Researchers have also developed

adaptations of the RFS which may be clinically useful including the Therapist Relationship Interview (TRI; Safran & Muran, 2007) and the Panic Specific Reflective Functioning Scale (PSRFS; Rudden et al., 2006). In addition to the adaptations of the RFS for adults, this rating system has also been modified for use with children and adolescents (e.g., the Child Reflective Functioning Scale (CRFS; Ensink, 2004) and Reflective Function Scale-Adolescent (RFS-A; Chow et al., 2014)).

The Mentalization Questionnaire

Development and Factor Structure

The RFS (Fonagy et al., 1998) reigned as the main measurement tool for assessing individuals' mentalization capacities for over a decade and was widely employed across various research domains. However, the laborious and time intensive nature of this tool necessitated the need for more efficient measures of mentalization that could be used to assess treatment-related change (Hausberg et al., 2012). To meet this need, Hausberg et al., (2012) developed the first self-report measure of mentalization, the Mentalization Questionnaire (MZQ), using a multi-step, peer reviewed process with a German patient sample. The care and complexity of their approach is atypical and merits some review. In the first step, patient statements were generated based on a review of literature in mentalization and psychopathology as well as the German translation of the manual for the Reflective Self Function Scale (Daudert, 2002), which is a precursor to the RFS (Fonagy et al., 1996). Items were then examined by an expert in psychological diagnostics and four patients undergoing psychotherapy. After suggested changes were made, the statements were reviewed by three other experts in Mentalization-Based Treatment (MBT; Bateman & Fonagy, 2004a), which then yielded 40 items. Subsequently, they conducted an item analysis in which 12 items were deleted due to negative item-total correlations, signifying that these items

did not discriminate between individuals with lesser and greater capacities for mentalizing. The final questionnaire consists of 15 items.

To examine the psychometric properties of the MZQ, Hausberg et al. (2012) administered the MZQ to two samples; the first sample included 97 inpatients in a psychiatric hospital while the second sample was composed of 337 inpatients in a psychotherapeutic setting. The MZQ was administered to the first sample (N= 97) to examine factor structure; exploratory PCA revealed a four-factor solution that explained 59% of the variance. The four factors are as follows: 1) refusing self-reflection, which encompasses avoiding thinking about mental states or an outright rejection of, and fear of being overwhelmed by affective experiences, 2) emotional awareness, which includes deficits in perceiving, identifying, and discriminating between diverse mental states, 3) psychic equivalence (one of three modes of primitive mentalization reviewed above), and 4) regulation of affect, which includes difficulties modulating emotional experiences and may lead to feelings of fear and helplessness. Factor loadings ranged from .57 - .68. In the final questionnaire, four items loaded onto each of the first three factors and three items loaded onto the last factor, regulation of affect. The internal consistency for the four factors ranged from $.58 < \alpha < .71$ and internal consistency for the final score was $\alpha = .81$. Adequate temporal stability was demonstrated ($.60 < r < .68$ for the individual factor scores and $r = .76$ for the total score), however the amount of time in between test administrations was not specified. Construct validity was demonstrated as patients who reported suicide attempts and self-harm behaviors had significantly lower mentalization scores than those without attempts or self-injurious behaviors. Furthermore, patients with BPD also scored significantly lower on the MZQ than patients with other disorders, and individuals that were classified as having secure attachment styles had significantly higher mentalization scores than insecurely attached individuals. It's worth noting

that the group sizes for all of these *t*-tests were unequal and some group sizes were quite small (*n* = 14 for the group with BPD and *n* = 18 for the securely attached group). Although Hausberg et al. (2012) did note that the small sizes may limit the usefulness of these findings, they did not discuss whether the comparison groups (e.g., securely vs. insecurely attached) had unequal variances. If this were the case, a Welch's *t*-test would have been more suitable for investigating differences between these groups and it is possible that the current findings, which reveal significant group differences, may represent a Type I error. Nonetheless, in support of construct validity, moderate negative correlations were found between MZQ scores and two measures of symptom severity ($-.51 < r < -.64$) as well as a measure of loss of identity ($r = -.50$), a core diagnostic feature of BPD. Hausberg et al. (2012) also found significant improvements in all MZQ scores from the time of admission to 6 months following psychotherapy treatment.

Other studies have found significant improvements in mentalization throughout a multifaceted treatment (including individual and group psychotherapy, psychoeducation, etc.) and a short-term mentalization based treatment for individuals with personality disorders (McGowan et al., 2021). McGowan et al.'s (2021) hierarchical linear regression model indicated that the change in mentalization scores from pre to post treatment was a significant predictor of the positive change in emotional reactivity and social functioning. Similarly, improvements in mentalization as measured by the MZQ have been found to predict decreases in reported interpersonal issues (Hayden et al., 2018). As such, these findings support construct and predictive validity of the MZQ as previous research which used the RFS (Fonagy et al., 1998) and the Reflective Functioning Questionnaire (RFQ; Fonagy et al., 2016) also found improvements in mentalization during various interventions (Belvederi Murri et al., 2017; De Meulemeester et al., 2018; Fischer Kern et al., 2015; Levy et al., 2006) and these improvements

were linked to positive changes in symptomatology (Chiesa et al., 2021; De Meulemeester et al., 2018; Müller et al., 2006).

Beyond Initial Validation

Hausberg et al. (2012) specified important limitations of this seminal validation study; they found slight differences in internal reliability between samples and thus, recommended re-examination of the factor structure in a larger, heterogenous sample. They also recommended research into divergent as well as convergent validity (i.e., associations with other measures of mentalization, such as the RFS). To address these limitations, five additional research teams have performed variations of factor analyses with European populations (Eloranta et al., 2020; Innamorati et al., 2017; Paridaens, 2012; Ponti et al., 2019; Raimondi et al., 2021). Four out of these five studies revealed that a unidimensional structure was the best fitting model, suggesting that the MZQ may capture one's general capacity for mentalizing (Innamorati et al., 2017; Paridaens, 2012; Ponti et al., 2019; Raimondi et al., 2021). Factor loadings ranged from .30 – .66 across these studies. In contrast to these findings, Eloranta et al., (2020) found that an alternative four-factor structure was a good fit for the data. This model was based on the original four subscales however, they moved six items to different factors. Reliability, convergent validity for the four factors, and loadings for each item onto their respective factor were adequate. In sum, four of the five validation studies were inconsistent with Hausberg et al.'s (2012) four-factor model and this may be explained by the use of non-clinical samples from the general population in these studies. Indeed, Hausberg et al. (2012) developed this scale for use with patients with a broad range of psychological disorders as a way to monitor their progress throughout psychotherapeutic interventions (MBT in particular). Thus, it is possible that groups of items may cluster together in different ways for clinical vs. non-clinical samples. For instance, items

measuring psychic equivalence are based on common ways that borderline individuals experience deficits in mentalization, as specified by Fonagy et al. (1996), however, individuals in the general population may experience different manifestations of lower mentalizing capacities. Moreover, both Eloranta et al. (2020) and Ponti et al. (2019) recruited adolescents and it is possible that the MZQ has a different factor structure for young adults due to their incomplete social, emotional, and psychological development. Inconsistent with this developmental hypothesis, the mean age (M = 40) for participants in Paridaens' (2012) study was similar to the initial validation study yet they still found an alternative factor structure to Hausberg et al. (2012). Furthermore, four of the five subsequent studies (Eloranta et al., 2020; Innamorati et al., 2017; Ponti et al., 2019; Raimondi et al., 2021) were translated into alternative languages (Italian and Finnish) and it is possible that semantic differences across languages may account for variation in factor structure. Other studies have examined the factor structure of other translated versions of the MZQ however, they were excluded from this review because the published research is written in Korean (Song & Choi, 2017) and Farsi (Drogar et al., 2020).

While the factor structure of the MZQ remains uncertain, construct validity has been demonstrated across a few studies. When compared to control groups without any psychological diagnoses, individuals who were diagnosed with an eating disorder (Ponti et al., 2019) and BPD (Vijayaraghavan et al., 2018), scored significantly lower on the MZQ. Moreover, mentalization was negatively related to severity of depression symptoms (r = -.68) and traumatic experiences in childhood (r = - .30) (Belvederi Murri et al., 2017; Zybutz et al., 2021). This latter finding is in line with theoretical work postulated by Fonagy and colleagues (Fonagy & Higgit, 1989; Fonagy et al., 1996). In their pioneering work, it is asserted that individuals who have experienced abuse perpetrated by their caregiver inhibit mentalizing others' internal states as a

self-protective defense that prevents them from considering their caregiver's ill-natured intentions. Relatedly, Hayden et al. (2018) found that scores on the MZQ were negatively associated with attachment anxiety (r = -.69) and attachment avoidance (r = -.57). These findings align with conceptual work that has outlined the importance of a secure relationship with one's caregiver for the development of mentalization capacities in early infancy and childhood (Fonagy et al., 1998). Moreover, Paridaens (2012) found a negative correlation with alexithymia (r = -.58), a construct which encompasses difficulties describing feelings and favouring externally oriented thoughts. Moderate correlations were also found with adaptive emotion regulation (r = .34), mindfulness (r = .42), and with the perspective taking (r = .29) subscale (which is purported to underly a cognitive empathy factor) on the Davis Empathy Scale (DES; Davis, 1980) (Paridaens, 2012; Schwarzer et al., 2021b). To the authors' surprise, the personal distress subscale (which supposedly underlies affective empathy and involves feeling anxious or fearful when observing others in distress) was negatively correlated (r = -.31) with the MZQ. Paridaens (2012) explains that mentalization may be more closely associated with cognitive, as opposed to affective empathy, however, this explanation is inconsistent with Luyten et al.'s (2020) assertions that mentalization involves both cognitive and affective content and processes. An alternative explanation for this finding is that the DES may not yield an accurate assessment of affective empathy as affectively laden items (e.g., "In emergency situations, I feel apprehensive and ill-at-ease") require individuals to use perspective taking (Chrysikou & Thompson, 2016). While the questionable validity of the DES makes it difficult to interpret the correlations between its subscales and the MZQ, the DES is an adequate measure of overall empathy. Thus, the lack of association between the overall score on the DES and the MZQ is

surprising given theoretical accounts which explain the overlap between these constructs (see *Mentalization as an Umbrella Term*).

Additional work on the MZQ revealed similar levels of internal consistency to Hausberg et al.'s (2012) validation study. Cronbach's α estimates range from .75 - .83 (Belvederi Murri et al., 2017; Innamorati et al., 2017; Paridaens, 2012; Ponti et al., 2019). Providing some initial evidence for convergent validity, the MZQ was significantly correlated (r = .24) with the MASC (Dziobek et al., 2006), a task-based measure of mentalization, in one study (Schwarzer et al., 2021a). Similarly, scores on the Reading the Mind from the Eyes Test (RMET; Baron-Cohen et al., 2001a), another task-based assessment of theory of mind/mentalization, predicted the refusing self-reflection and emotional awareness factors on the MZQ in a clinical sample (Fekete et al., 2020). It is worth noting that the RMET is a measure of sensitivity to facial expressions in which emotions are displayed and performance on this measure is ostensibly related to identifying and processing others' mental states and emotional experiences (Baron-Cohen et al., 2001a). This measure has been criticized given that there are two underlying assumptions which remain unresolved, 1) whether performance is indicative of accuracy in identifying psychological states (as images were taken from magazines and did not have an indication of emotional experience of the target) and 2) whether specific psychological states (e.g., "preoccupied") can be inferred from one's eyes (Johnston et al., 2008). Moreover, Oakley et al. (2016) found that alexithymia but not severity of autism spectrum disorder symptoms significantly predicted performance on the RMET. They concluded that the RMET assesses emotion recognition as opposed to theory of mind capacities or understanding others' mental states. Taken together, it is debatable whether the RMET is indicative of mentalization.

Critiques and Future Directions

From a qualitative perspective and through examining face validity of the MZQ, there are important considerations regarding both the item content as well as the original factor structure that merit discussion. First, it is stated that two of the three primitive modes of mentalization (psychic equivalence, teleological, and pretend mode) are captured in this scale (Fonagy & Target, 1996). Hausberg et al. (2012) and Rishede et al. (2021) state that the refusing self-reflection scale captures elements pertaining to the teleological mode, in which one's understanding of mental states is based on concrete, physical evidence. For instance, in the teleological mode, an individual makes inferences about a person's intentions based on seemingly unambiguous actions rather than perspective taking. While the following item is a good demonstration of the teleological mode, "If someone yawns in my presence, that's a reliable sign that he is bored in my company", it is debatable whether individuals experiencing the teleological mode would endorse other items on this subscale that reflect a fear of experiencing emotions (e.g., "Most of the time it is better not to feel anything" and "Talking about feelings would mean that they become more and more powerful") because they often do not even recognize internal states unless related to some tangible goal (Fonagy et al., 2015). The psychic equivalence scale is proposed as the third factor. In this mode individuals equate their inner experiences with external reality (e.g., expectations of threats are experienced with a similar level of distress as external events that are actually threatening). Although about half of the items in this scale seem to be good indicators of the psychic equivalence mode (e.g., "If I expect to be criticized or offended, my fear increases more and more"), another item ("I only believe that someone really likes me a lot if I have enough realistic proof for it (e.g., a date, a gift or a hug)") seems to align better with the teleological mode of thinking.

In addition to the theoretical and empirical work surrounding the primitive modes of mentalization, Luyten et al. (2020) proposed that mentalization encompasses four dimensions: implicit/explicit or automatic/controlled mentalizing, cognitive and affective content and processes of mentalizing, mentalizing about oneself and others, and mentalizing based on internal or external cues. The MZQ captures some of these dimensions but lacks representation of others. For instance, items cover content pertaining to an individuals' reflections on their own as well as others' mental states (although there are many more items covering content related to the self). Regarding individuals' appraisals of others' mental states, inferences may be implicit and automatic or explicit, controlled, and deliberate. It seems that the explicit/controlled typology is represented in the MZQ as most references to others' mental states involve a logical and effortful thought process to reach a conclusion (such as interpreting a yawn as a sign of boredom or that someone is fond of you because they give you gifts). While there are no items that capture quicker, less intentional inferences about others' mental states, it is debatable whether automatic mentalizing can be captured in a self-report measure and whether task-/experiment-based measures may be better suited to assessing this component in real time. Affective processes of mentalization (specifically emotional contagion) are not represented but cognitive, perspective-taking features of mentalization are captured. For instance, a yawn is interpreted as a sign of boredom because individuals in the teleological mode have a self-centered outlook on things and fail to think about others' perspectives (Hausberg et al., 2012). Last, inferences about others are clearly oriented toward external cues (such as one's actions) with little reference to their internal states. Overall, the lack of representation of the emotional contagion and internal dimensions of mentalization, and the over-emphasis on self-related

mentalizing are weaknesses of the MZQ. It is also debatable whether 3-4 items can adequately capture the different forms of impaired mentalizing that are put forward in this scale.

After evaluating research on the MZQ it is clear that there is abundant support for construct validity and preliminary evidence for convergent validity and internal consistency. Future research should continue to explore the factor structure of the English version of the MZQ with clinical samples to gain clarity on the discrepant findings discussed here.

The Reflective Functioning Questionnaire

Development and Factor Structure

Following the development of the MZQ, Peter Fonagy and his colleagues further extended their work on mentalization by developing the Reflective Functioning Questionnaire (RFQ; Fonagy et al., 2016), a self-report questionnaire based on the RFS. The authors sought to develop a brief screening tool that could be useful for epidemiological studies in which researchers could investigate the role of impaired mentalizing in the etiology of personality disorders and in environments that foster the development of insecure attachment styles. They set out to capture two types of mentalizing impairments in their measure, namely *hypomentalizing*, which reflects a severe lack of mentalization skills (e.g., unable to develop complex models of mental states), and *hypermentalizing* (or pseudomentalizing), which is characterized by articulate explanations of mental states that are baseless and reflect inaccurate models of others' and one's own mind. To develop the RFQ the authors generated 101 potential items, including polar response and central responses items, both of which were rated on a Likert scale ranging from strongly disagree (1) to strongly agree (6). Neither the process through which these items were generated nor the sources of the content were specified in the first publication of the measure (Fonagy et al., 2016). Subsequently these 101 items were rated by 14 mentalization experts and

after two rounds of reducing items (e.g., removing statements that lacked reliability and were redundant) 46 items remained. After PCA and CFA with a community and clinical sample (composed of outpatients with BPD and eating disorders), construct validity of both the polar and central response items was not established. As such, they recoded the 26 central response items and developed two subscales which correspond to the two types of impairments in mentalizing: Certainty about Mental States (RFQ_C) referring to overconfidence in assessing one's own and other's mental states, and Uncertainty about Mental States (RFQ_U) which refers to extreme difficulties in understanding others' mental states. The final version included 8 items, but the author team also developed a 46 and 54 item version in tandem with the shortened version. Most research with the RFQ has used the 8-item version (Müller et al., 2021). The RFQ_C includes items 1-6 while the RFQ_U includes all items except 1 and 3 (e.g., item #4 "When I get angry I say things that I later regret" is included on both scales). For the RFQ_C scale, responses on the lower end of the Likert scale indicating strong disagreement were given the highest scores, signifying that an individual is overly certain of mental states. For the RFQ_U scale, responses towards the higher end of the scale indicating strong agreement were given the highest scores, which are indicative of a severe lack of mentalizing.

The first study using this scale included 295 healthy controls, which included university students and staff as well as 108 outpatients with BPD and/or eating disorders. CFA indicated a two-factor structure had a satisfactory model fit across the samples. Factor loadings for both scales were higher in the clinical sample with a few exceptions. The subscales were moderately correlated with one another (r = -.61 for the clinical sample and r = -.33 for the control sample) and internal consistency of the subscales was adequate (α ranged from .65 – .77 in the clinical sample and .63 – .67 for the controls). To examine test-retest reliability, 50 participants (30

healthy controls and 20 outpatients) completed the RFQ three weeks after the initial administration. Test-retest reliability was excellent (r = .84 for RFQ_U and r = .75 for RFQ_C). Using *t*-tests, both subscales discriminated between controls and outpatients, yet in a regression analysis, only the RFQ_U significantly predicted BPD diagnosis. In the clinical sample, the RFQ_C was positively associated (r = .30) with measures of empathy but not significantly related (r = .14) to scores on the Reading the Mind in the Eyes Test (RMET) which the authors refer to as a measure of "externally based mentalizing" (Fonagy et al., 2016, p. 5). The RFQ_U was not significantly correlated with any of these measures (r =-.05 with empathy r = .03 with the RMET). As noted above, many researchers are dubious as to whether the RMET is indeed a measure of mentalization.

A second study by Fonagy et al. (2016) with a clinical sample (N = 129) similarly found that the RFQ_U was highly associated with having a diagnosis of BPD (OR = 1.31, CI = 1.12 - 1.53) as well as elevated levels of self-harm, depressive symptom severity, social impairment, and issues with anger control (.33 < r < .40). In contrast, higher scores on the RFQ_C were not associated with BPD diagnosis (OR = .94, CI = .77 - 1.16) and correlations with psychopathology variables (e.g., with self-harm, depressive symptom severity, and social impairment) were negative (-.09 < r < -.27). These null associations suggest that being overly certain of one's own and others' mental states is associated neither with symptom severity nor degree of functional impairment. These findings are in contrast to the theoretical assertion and preliminary empirical evidence suggesting that *hypermentalizing* plays a role in the etiology and/or maintenance of psychopathology (Fonagy & Luyten, 2016; Fonagy et al., 2016; Sharp et al., 2011). To examine convergent validity of the RFQ, the authors conducted a third study and examined the associations between parents' scores on the RFQ and Parental Reflective

Functioning Questionnaire (PRFQ; Rutherford et al., 2015), with their infants' attachment security as measured by the Strange Situation Procedure (SSP; Ainsworth et al., 1978). The RFQ_C and RFQ_U were moderately correlated with the 1) Prementalizing Modes subscale ($|.27| < r < |.29|$) and 2) Certainty about Mental States subscale ($|.27| < r < |.41|$), but not the Interest and Curiosity in Mental States subscale ($.013 < r < .024$) of the PRFQ, suggesting that these measures are related but represent distinct capacities (e.g., overall mentalizing versus understanding one's children's mental states). While the two RFQ subscales were only weakly correlated with infant attachment insecurity (r = -.16 with the RFQ_C and r = .11 with the RFQ_U), the RFQ_C (OR = 1.30) but not the RFQ_U significantly predicted infant attachment insecurity in a binary regression analysis. Based on this set of seminal studies, many findings support construct validity (e.g., moderate associations between psychopathology variables and the RFQ_U) and convergent validity (e.g., moderate correlations between the RFQ and PRFQ), yet some results are inconsistent with underlying theoretical assumptions (e.g., weak relations between infants' attachment and parents' performance on the RFQ and between indicators of psychopathology and the RFQ_C).

Beyond Initial Validation

Among all the self-report measures which assess mentalization the RFQ has been most widely adopted as ~400 papers have cited the seminal work by Fonagy and colleagues (2016). Amidst this extensive work utilizing the RFQ, research has found evidence supporting the 2-factor structure, construct validity, and adequate internal consistency of this instrument. Nine studies (Badoud et al., 2015; Bizzi et al., 2021; Cosenza et al., 2018; Griva et al., 2020; Morandotti et al., 2018; Mousavi et al., 2021; Müller et al., 2021; Spitzer et al. 2020; Wozniak-Prus et al., 2022) have re-examined the factor structure of the RFQ with only one of these studies

investigating the psychometric properties of the English version (Wozniak-Prus et al., 2022). These studies primarily recruited non-clinical samples of adolescents and adults with the exception of two projects (Morandotti et al., 2018; Müller et al., 2021). The results of six studies (Badoud et al., 2015; Bizzi et al., 2021; Cosenza et al., 2018; Griva et al., 2020; Morandotti et al., 2018; Mousavi et al., 2021) confirmed the 2-dimensional factor structure composed of the RFQ_C and RFQ_U. Conversely, the remaining three studies concluded that the best fit model has a unidimensional structure (Müller et al., 2021; Spitzer et al., 2020; Wozniak-Prus et al., 2022). In Wozniak-Prus et al.'s (2022) study, all items had adequate factor loadings onto the single factor (ranging from .35 – .81) with the exception of item 1 which had a lower factor loading of .19. In contrast to the other loadings, item #7 negatively loaded onto the unitary factor because this item captures certainty about mental states whereas all other items represent uncertainty. Müller et al. (2021) highlighted many psychometric issues resulting from the fact that four of the eight items are double scored to calculate the RFQ_U and RFQ_C. Specifically, given that the rescaled scores for each of the subscales are based on a single rating provided by the respondent, these scores are dependent on one another and provide redundant information. To further strengthen this point, they assert that polychoric correlations between the subscales (which account for the ordinal level scale) would yield a perfect negative correlation (r = -1) and maintained that past validation studies did not encounter this issue in CFA because they treated the subscales as continuous variables and applied robust maximum likelihood estimation. In view of these issues, the authors used an alternate scoring procedure to generate a single score and found that the unidimensional model provided the best fit to the data They argue that hypo- and hyper-mentalizing should be represented as two maladaptive poles on one dimension with intermediate scores representing the most adaptive form of mentalizing. They also conducted

CFA on the data using the original double-scoring method and results indicated that the 2-dimensional structure was the best fit for the data in this form.

Internal consistency estimates similar to the original study have been found by researchers who translated versions of the RFQ for other populations (Badoud et al., 2015; Griva et al., 2020; Morandotti et al., 2018; Mousavi et al., 2021). Across this work, α ranged from .71 – .80 for the RFQ_C and .62 – .79 for the RFQ_U. Mean inter-item correlations were found to range from .28-.36 (Badoud et al., 2015; Morandotti et al., 2018) and test-retest reliability was *r* = .81 for the RFQ_C and ranged from *r* = .78 – .85 for the RFQ_U (Morandotti et al., 2018; Mousavi et al., 2021). The time interval between testing was 2 weeks in Morandotti et al.'s (2018) study and 7 weeks in Mousavi et al.'s (2021) work. Research using the English version of the RFQ has found similar psychometric properties (Li et al., 2020; Sacchetti et al., 2019).

Lending support to construct validity, various studies have found that mentalization, as measured by the RFQ, partially mediated the association between childhood maltreatment, physical, and emotional abuse and various mental health difficulties in adulthood (e.g., depressive and PTSD symptoms, self-harm and suicidal thoughts and attempts, dissociative experiences, and aggressive behavior; Berthelot et al., 2019; Huang et al., 2020; Li et al., 2020, Schwarzer et al., 2021b; Stagaki et al., 2022). In some of these studies, both hypomentalizing (measured by the RFQ_U) and hypermentalizing (measured by the RFQ_C) were tested as mediators for the above associations (Li et al., 2020; Schwarzer et al., 2021b). Both subscales were significant mediators in Li and colleagues' (2020) study, however only the RFQ_C emerged as a partial mediator between emotional abuse and aggressive behavior in adulthood in Schwarzer et al.'s (2021b) work. In the remaining studies mentioned above a latent variable representing mentalization was generated based on the two subscales of the RFQ. These findings

are consistent with theoretical tenets regarding the development of mentalization, specifically the notion that childhood abuse confers vulnerability to later psychopathology in part due to the development of impaired mentalizing capacities (Fonagy & Target, 1997). Parents who perpetuate abuse/maltreatment form impaired attachments with their children as they fail to respond sensitively to their children's needs and affect. As such, children do not have the opportunity to explore, reflect on, and understand their own mental states within the safety of a close, intimate relationship. In support of this notion, Handeland and Kristiansen (2017) found that high scores on the RFQ_U (representing a markedly low capacity for mentalizing) were two to three times more common in mothers who reported many instances of trauma in adolescence versus those who reported far fewer instances of trauma in their early life. It is also postulated that individuals who have experienced trauma may deliberately inhibit their mentalization capacities to avoid distressing affective states that are both associated with the trauma and would result from considering the perpetrator's malevolent intentions (Choi-Kain & Gunderson, 2008). Although this approach may serve a protective function in the short-term, continuously avoiding attending to others' mental states may lead to feelings of isolation and a lack of connection, which could prompt self-harm (Luyten et al., 2020). Additionally, recent work suggests that even when parents do not inflict abuse, parents of children who have experienced abuse have lower parental reflective functioning than parents of non-abused children (Ensink et al., 2017). This finding may suggest that parents of abused children do not foster an adequate environment for children to develop mentalization capacities, whether parents inflict the abuse or not. Last, Berthelot et al. (2019) briefly discuss a pathway in which early abuse leads to changes in neurobiological processes, which may influence the development of mentalizing.

In contrast to the above research, some empirical work has failed to support construct validity of this tool (in particular the RFQ_C). According to mentalization theory, one's capacity for mentalization develops within the context of secure attachment relationships in early childhood and extensions of this theory suggest that impaired mentalization should be associated with patterns of insecure attachment. In line with these assumptions, multiple studies have found that the RFQ_U is positively associated with both anxious ($.25 < r < .66$, $p < .05$) and avoidant attachment ($.19 < r < .38$, $p < .01$ for four of five analyses). In contrast, the associations between the RFQ_C with both anxious ($-.61 < r < -.22$, $p < .05$ for four of five analyses) and avoidant attachment ($-.37 < r < .09$, $p < .01$ for four of five analyses) are of similar strength yet they are negatively correlated (Barberis et al., 2022; Brugnera et al., 2021; Green et al., 2021; Huang et al., 2020; Stagaki et al., 2022). A similar pattern of associations has emerged in research examining the relations between the RFQ and psychopathology and between the RFQ and alexithymia, a construct which captures difficulties in labeling, describing, and focusing on emotions. First, a number of studies have found that while the RFQ_U was positively related to PTSD, BPD, depressive, anxiety, and eating disorder symptoms, the RFQ_C was negatively associated with these symptoms (Barberis et al. 2022; Berthelot et al., 2019; Bizzi et al., 2021; Huang et al., 2020; Kahya & Munguldar, 2022; Li et al., 2020; Quattropani et al., 2022; Vahidi et al., 2021). Moreover, two studies found that non-clinical control groups scored significantly higher than individuals with eating disorders on the RFQ_C whereas this pattern was reversed for the RFQ_U with clinical populations scoring significantly higher than controls (Cucchi et al., 2018; Sacchetti et al., 2019). Second, a few studies have found that alexithymia is significantly positively related to the RFQ_U ($.13 < r < .66$, $p < .05$) but negatively associated with the RFQ_C

(-.55 < r < -.21, p<.05) (Badoud et al., 2015; Barberis et al., 2022; Bizzi et al., Calaresi & Barberis, 2019; Cucchi et al., 2018; Morandotti et al., 2018; Mousavi et al., 2021).

Taken together, these results suggest that being highly uncertain about mental states (e.g., hypomentalizing) is related to insecure patterns of adult attachment and increasing levels of psychopathology and alexithymia, which aligns with mentalization theory and findings in prior research with other measures, namely the RFS (Ekeblad et al., 2016; Fischer-Kern et al., 2010; Gullestad et al., 2013; Levy et al., 2006; MacBeth et al., 2011; Ward et al., 2001), MentS (Benoit, 2020; Dimitrijević et al., 2018), MZQ (Hausberg et al., 2012; Hayden et al., 2018) and MMQ (Gori et al., 2021). In contrast, higher certainty about mental states (e.g., hypermentalizing) was found to relate to lower levels of insecure attachment, psychopathology, and alexithymia. These latter findings are inconsistent with the theoretical assertions specifying that hypermentalizing also represents a deficit in mentalizing (Fonagy et al., 2016) and that impaired mentalizing develops within the context of an insecure attachment style as a child and confers vulnerability to the development and maintenance of psychopathology (Luyten et al., 2021). It is also noteworthy that these findings on the RFQ_C are contrary to prior research examining hypermentalizing using the MASC. Specifically, a few studies have found that hypermentalizing on the MASC predicts symptoms of numerous psychological disorders (Abate et al., 2017; Li et al., 2020; Sharp et al., 2011) and is related to insecure attachment style in children and adults (Cortés-García et al., 2021; Henry et al., 2022). Moreover, a recent meta-analysis concluded that there was sufficient evidence to support the association between hypermentalizing (when measured with the MASC) and psychopathology (r = .24, 95% CI = .17 – .31) and that neither age nor gender moderated this association (McLaren et al., 2022). These contrary results regarding hypermentalization and psychological difficulties may suggest that the

MASC and the RFQ are measuring related but not identical social-cognitive constructs. Perhaps the MASC is better capturing extreme forms of hypermentalizing which represent a marked deficit in mentalizing. In contrast, the RFQ_C may be a more suitable measurement tool for capturing some degree of overconfidence in making inferences about others' behaviors and accompanying mental states, but not at a level which would cause social impairment.

In view of the above contradictory findings regarding hypermentalization, it is particularly useful to investigate whether the RFQ is related to other measures of mentalization. To the authors' knowledge, only three studies (Anis et al., 2020; Handeland et al., 2019; Raimondi et al., 2021) have examined the convergent validity of the RFQ. Anis et al. (2020) recruited parents and investigated the associations between the RFQ, PRFQ, and the RFS based on the Parent Development Interview (PDI; Slade et al., 2004). Neither RFQ subscale was significantly correlated with the RFS rated PDI self-score (e.g., parent), child score, or total score (r = -.032, p = .70 for RFQ_C and the total score and r = .026, p = .75 for the RFQ_U and the total score). The authors stated that they expected these null findings given the potential differences in the capacities each measure assesses. While the RFQ is a measure of one's general mentalizing and is not embedded with a particular context, the PDI rated RFS is meant to elicit mentalizing within the specific context of a parent-child relationship. However, even when considering the differences in scope of the construct it is reasonable to assume a moderate association between a general capacity for mentalizing and an ability to consider and reflect on children's mental states. Indeed, the seminal research by Fonagy et al. (2016) found significant correlations between the RFQ_U and RFQ_C with two of three subscales of the PRFQ (-.26< r <.41, p<.04), which is a self-report questionnaire. Moreover, Raimondi et al. (2021) found that an Italian adaptation of the MZQ was significantly related to both the RFQ_U (r = -.41, p<.001)

and the RFQ_C (r =.60, p<.001). Given these inconsistent findings, perhaps the above null association may be explained by the differences between self-report and interview-based measures. Self-report questionnaires require the participant to have insight into their own skills and may be more vulnerable to biased forms of reporting whereas the interviewer's capacity to discern different levels of mentalizing plays a crucial role in interview-based assessments.

Contrary to Anis and colleagues' (2020) results, Handeland et al. (2019) found a significant association between PDI rated RFS scores and the RFQ_U (r = -.52, p <.01) but not the RFQ_C (r = .12, p > .05). Anis et al. (2020) suggest that these findings are inconsistent because Handeland et al. (2019) used a new, Norwegian version of the RFQ in which clinical cut off scores were not yet validated. They also implied that the association between the RFQ_U and the PDI rated RFS is spurious because a corresponding significant association between the RFQ_C and the RFS rated PDI would be expected given strong association between the RFQ_U and RFQ_C (r = -.60). Conversely, Handeland et al. (2019) suggest alternative reasons for the null association between the RFQ_C and the PDI rated RFS. They assert that in the context of the strong negative association between the subscales it is difficult to obtain high scores on both scales and given that most participants scored highly on the RFQ_U, fewer scored high on the RFQ_C. This led to a low mean score and less variability in the distribution of the RFQ_C scores, resulting in reduced power and primarily medium level scores which were indicative of balanced mentalizing. As such, the null association was attributed to the strongly associated subscales (which is likely a result of double scoring most items) and resulting statistical repercussions mentioned above.

Critiques and Future Directions

As previously mentioned, there are two main theoretical frameworks in which mentalization capacities are explained. Luyten et al. (2020) put forward a multidimensional model where they highlighted 4 dimensions or polarities through which mentalizing can occur. Complementing this framework, researchers have introduced the dual deficit model, referring to hypomentalizing and hypermentalizing (Fonagy & Luyten, 2016; Fonagy et al., 2016; Sharp et al., 2011). Given the little evidence accumulated for convergent validity of the RFQ_C, it appears that this tool only partially aligns with the dual deficit framework. The following section will further examine whether the RFQ maps onto the multidimensional framework.

After examining all items on the RFQ, it appears that only a few polarities on the above dimensions are captured, including understanding one's own and others' mental states. While one's capacity for understanding their own (e.g., item #2, "I don't always know why I do what I do') and others' mental states (item #1, "People's thoughts are a mystery to me") is represented on the RFQ, there is a disproportionate number of items assessing mentalizing about oneself (e.g., seven out of eight items), which may suggest that the RFQ fails to adequately capture the one's capacity to understand others. Neither cognitively nor affectively based mentalizing appear to be represented on the RFQ as the only item referring to understanding others' beliefs, desires, and emotions, is broad (e.g., item #1) and lacks precision. Particular examples of cognitive processing involved in mentalization are perspective taking and belief-desire reasoning (e.g., ability to predict an individual's actions through an understanding of their beliefs and desires) yet neither of these were captured in the single item reflecting mentalizing about others. While item # 8, "strong feelings often cloud my thinking" does refer to affect, the phenomenon of emotional contagion is not captured given that this item does not reference others. However, it is worth noting that Luyten et al. (2020) do not clearly define how affective processing should be

operationalized in relation to mentalizing about one's *own* internal states. Last, none of the items appear to capture *controlled* or *automatic* mentalizing, although this latter form of mentalizing may be related to the various items which seem to capture impulsivity (e.g., "When I get angry I say things that I later regret"). Aside from the lack of correspondence between the RFQ and the multidimensional model, there are other practical limitations to this measure. For instance, items 3 and 4 appear to capture redundant information about emotion regulation, which is especially problematic given the brief nature of this measure. Moreover, Müller et al. (2021) assert that one's inclination towards impulsivity during periods of negative affect may not necessarily be related to a lower capacity to mentalize and could instead be attributed to poor stress reactivity. As such, these items may be measuring another construct that may be contained within the boundaries of mentalization. Another potential issue is that seven of eight items are framed such that they capture difficulties with mentalizing (e.g., "I don't always know why I do what I do"), signifying that high scores on the RFQ_C are indicative of a strong rejection of items tapping uncertainty, rather than endorsing items which specify certainty about mental states. As a result, the RFQ_C may not capture extreme deficits in hypermentalization which could explain the negative associations with insecure attachment, psychopathology, etc.

After examining a myriad of research using the RFQ, it appears there is abundant support for construct validity of the RFQ_U as an indicator of markedly impaired mentalizing and that this deficit is related to insecure attachment, prior traumatic experiences, alexithymia, and psychopathology. Yet it remains uncertain whether the RFQ_C sufficiently assesses one's tendency to hypermentalize or overinterpret mental states to the extent that it causes impairment. Additionally, prior research provides adequate support for internal consistency and reliability of this measure, however the dimensional structure and scoring procedures recommended are

contentious. Last, whether the RFQ and particularly, the RFQ_C has sufficient convergent validity with other measures of mentalizing remains questionable and merits further investigation.

The Mentalization Scale

Development and Factor Structure

With the growing interest in self-report measures of mentalization, Dimitrijević et al. (2018) sought to develop a comprehensive scale which closely adheres to the original conceptualization of mentalization, the Mentalization Scale (MentS). This measure was initially developed in Serbian but later translated into English and other languages. Fifty-five items were generated from various sources including a Serbian measure of attachment (the Revised Questionnaire for Attachment Assessment; Hanak, 2004), the RFS manual (Fonagy et al., 1998), and the Handbook of Mentalization-Based Treatment (Allen, 2006) and were later reviewed by two experts in attachment and mentalization. The earliest iteration of the MentS was administered to university students majoring in psychology and education (N =102) who provided feedback on the clarity and format of the measure. Ambiguous items were either revised or deleted and items which had item-total correlations less than .30 were deleted. Items that would increase internal consistency (e.g., alpha values) if excluded, were also deleted. This process yielded two amended versions with 30 items and 37 items, respectively, which were later administered to two university student samples (N = 87 and N= 91). A similar pruning process was again employed, resulting in the final questionnaire composed of 28 items. In the first validation study with the final version of the MentS, Dimitrijević et al. (2018) recruited a large sample (N = 540) of workers at a dairy factory and university students and conducted PCA with an oblimin rotation using the eigenvalues greater than 1 guideline. This analysis indicated a

seven-factor model accounting for 55% of the variance. Dimitrijević et al. (2018) then used parallel analysis (O'Connor, 2000) and found that a three-factor model fit the data well, explaining ~38% of the variance. The factors include other-related mentalization (MentS-O), self-related mentalization (MentS-S), and motivation to mentalize (MentS-M), referring to one's "need to understand the psychic world of self and others" (Dimitrijević et al., 2018, p. 6). Ten items each loaded onto MentS-O and the MentS-M factors, while eight items loaded onto the MentS-S. Factor loadings for individual items with their corresponding factor, ranged from .30 – .75, with the MentS-S yielding the highest factor loadings. In terms of internal consistency, item-total correlations ranged from .26 – .51 and Cronbach's α was between .76 – .77 for each of the three subscales and α = .84 for the total MentS score. In support of construct validity, attachment avoidance was significantly negatively associated the three subscales (r = -.31 for all subscales) and total score (r = -.41) on the MentS. Attachment anxiety was negatively related with self-related mentalization (r = -.54), other-related mentalization (r = -.22), and the total score (r = -.32), but not motivation to mentalize. An ANOVA also revealed that individuals with a secure attachment style had significantly higher scores on the subscales and total-score than those with insecure attachments. Positive associations between the Empathy Quotient (Baron-Cohen & Wheelwright, 2004) and the MentS provide evidence of construct validity ($.35 < r < .51$). Similarly, skills pertaining to emotional intelligence (i.e., competencies in navigating and tolerating emotions) should theoretically be linked with mentalization capacities, and specifically, representing emotional mental states. While this relationship has seldom been investigated, Dimitrijević et al. (2018) found moderate positive associations between the MentS total score and subscales with a self-report measure of trait emotional intelligence ($.35 < r < .63$). Correlations between the MentS and a test of emotional intelligence where individuals are asked

to choose the most suitable response to a problem, were lower but still significant (.21 < *r* < .44). The moderate associations between empathy, emotional intelligence, and mentalization indicates a degree of conceptual overlap that is consistent with theoretical tenets but also suggests that there are important conceptual distinctions between these constructs. Last, Dimitrijević et al. (2018) were interested in the association between personality variables specified by the Big Five model and the MentS. Total and subscale scores were positively related to agreeableness (.08< *r* < .18), extraversion (.2 < *r* < .41), openness (.31 < *r* < .46), and conscientious (.35 < *r* < .48), and negatively associated with neuroticism (-.16 < *r* < -.53). The authors argued that it was sensible for this array of personality traits to be related to higher mentalizing skills. For instance, individuals that are more open to new experiences (high openness), socially outgoing (high extraversion), and have stable affect (low neuroticism) are likely motivated to understand both their own and others' mental states and emotional experiences. However, these findings are in contrast to Fonagy and colleagues' (1998) initial studies with the RFS where they did not find a significant relationship between the RFS and personality traits of extraversion and neuroticism.

Dimitrijević et al. (2018) aimed to expand the usefulness of this measure by validating it on a clinical sample. They included both inpatients and outpatients who had received a diagnosis of BPD (*n* = 62) as well as a nonclinical sample (*n* = 62). The groups were matched on age, gender, and education level. For the clinical group, Cronbach's α was similar to the first study for the MentS_S (α = .79) and MentS_O (α = .75) but lower for the MentS_M (α = .60) and the total score (α = .75). Similar to the first non-clinical sample, associations with both attachment avoidance (-.25 < *r* < -.53) and anxiety were negative (-.34< *r* < -.56). Yet correlations between attachment anxiety and the MentS_O (*r* = -.02) and with the MentS_M (*r* = -.06) were much lower. Relationships with personality variables were also of a similar strength to study 1, except

correlations with agreeableness were higher (.30 < *r* < .55). The control group had significantly higher scores on the MentS than BPD group. Last, when Dimitrijević et al. (2018) re-ran the parallel analysis with the clinical group, a three-factor structure was again specified however some items that were previously identified as loading onto one factor loaded onto a different factor. For instance, item 25, "I can easily describe what I feel" loaded onto the MentS_O in the non-clinical sample, but the MentS_S in the clinical sample. The authors explain that the emergence of slightly different factor structures may have been due to different sample sizes (*N* = 540 vs. *n* = 62). They refer to an unpublished German validation study (Dimitrijević et al., 2017) which replicated the initial factor model.

Beyond Initial Validation

Dimitrijević et al.'s (2018) study has been cited around 50 times and most of the subsequent work on the MentS has showcased construct and convergent validity as well as internal reliability. Only two studies have re-examined the factor model (Ahmadian & Ghamarani, 2021; Jańczak, 2021). Specifically, Jańczak's (2021) results supported the three-factor structure with the exception of item 15 on the MentS-M, which yielded a much lower factor loading. Ahmadian and Ghamarani (2021) used EFA on a Persian adaptation and this study garnered support for the three previously described factors as well. Reliability of these two adaptations was acceptable but Cronbach's α values were much higher across all factors for the Polish version (α ranged from .53 – .66 in Ahmadian and Ghamarani's (2021) study and .74 – .86 in Jańczak's (2021)). Some studies have found very similar alpha values to Dimitrijević et al.'s (2018) estimate based on a clinical sample. Djordjevic & Dordević (2019) found α ranged from .73 – .77 on all subscales except the Ments_M where α = .6. Gagnon's (2020) estimate was similar (α = .80), as was the estimate from Benoit (2020; α = .78), and Stanojević et al. (2020;

MentS_S, α = .79). In contrast, Richter et al. (2021) found lower Cronbach's α values (.55 – .70) in a small sample of inpatients in a psychiatric hospital (N = 26). Bhola and Mehrota (2021) found higher internal consistency values (.70 < α < .90) for a group of psychotherapists. Interestingly, the alpha value for the motivation to mentalize scale was notably lower than the other subscales and the total score across a few studies with clinical and community samples (Dimitrijević et al., 2018; Djordjevic & Dordević, 2019; Richter et al., 2021). This preliminary trend may suggest that the items within the motivation to mentalize scale do not reflect one single factor. When examining the face validity of this scale, it seems that most items are capturing either an individual's attribution of the degree of importance of mentalizing or the frequency with which an individual engages in mentalizing, both of which might be good indicators of an individual's motivation to reflect on mental states. However, it is possible that motivation to think about one's own versus others' mental states may be better represented as separate factors, rather than incorporating both types of motivation into one factor. Additionally, other items on this scale seem to reflect general interests that likely do not discriminate between individuals with higher versus lower mentalization capacities (e.g., "I like reading books and newspaper articles about psychological subjects") which might be decreasing the internal consistency of this scale and detracting from its construct validity.

In addition to work that has suggested an acceptable to moderate level of internal consistency, construct validity for the MentS has also been demonstrated. Stanojević et al. (2020) found negative correlations between self-related mentalization and self-report measures of anxiety (r = -.47) and depressive symptoms (r = -.35). The results of Francoeur et al.'s (2020) work also indicated that higher mentalization scores were associated with fewer psychiatric symptoms (r = -.21), as measured by the Brief Symptom Inventory (BSI; Derogatis, 1993).

Stanojević et al. (2020) reported a negative correlation (r = -.28) between the MentS_S and preoccupied attachment style, which typically arises when caregivers are inconsistent in their responses to children's cues and this attachment style is characterized by a positive model of others and a negative model of one's self (Bartholomew & Horowitz, 1991). In Djordjevic and Dordević's (2019) work, attachment anxiety was negatively correlated with self-related mentalization (r = -.40), but much weaker correlations were found with the other scales ($|.05| < r < |.16|$). Attachment avoidance was also negatively associated with all MentS scales ($-.18 < r < -.23$; Djordjevic & Dordević, 2019). Benoit (2020) found similar associations in a sample of early childcare teachers, however, in this case mentalization skills were more strongly associated with attachment avoidance (r = - .35) than attachment anxiety (r = .15). Taken together, these findings are consistent with theoretical work asserting that secure attachment relationships provide the groundwork for mentalization capacities and that impairments in mentalization are linked to psychopathology. Moreover, the MentS_S was shown to mediate the association between preoccupied attachment style and depressive symptomatology (Stanojević et al., 2020) and this finding is similar to a theoretical proposition that insecure attachment style brings about reduced capacities for mentalization which in turn leads to BPD (Bateman & Fonagy, 2004b).

To the authors' knowledge, only one study has investigated the convergent validity of the MentS with other measures of mentalization. Richter et al. (2021) examined the correspondence between the MentS and the RFS (Fonagy et al., 1998) coded from the Brief Reflective Function Interview (BRFI; Rudden et al., 2006). Scores on the MentS subscales and the RFS were moderately correlated ($.41 < r < .56$) and there was an even stronger association between the MentS total score and the RFS (r = .65), providing evidence for convergent validity.

Despite some promising work suggesting good convergent and construct validity of the MentS, other research has failed to support its construct validity. For instance, while theoretical work assumes an association between mentalization and recognizing emotional expressions, Djordjevic and Dordević (2019) failed to find significant correlations between one's ability to recognize various emotions (from facial expressions) and the total score on the MentS (Spearman's rho ranged from .03 - .14). Additionally, Gagnon (2020) found very weak associations between mentalization and attachment to maternal or paternal caregivers ($-.10 < r < .04$). Gagnon's (2020) results did indicate a negative association between mentalization and anxious attachment style to one's current romantic partner ($r = -.17$). Other research using the MentS_S has shown that therapists' mentalizing capacities regarding their own mental states is negatively associated ($-.20 < r < -.38$) with some maladaptive countertransference experiences. Yet, the MentS_M and MentS_O were largely unrelated ($.01 < r < .19$) to these countertransference experiences (Bhola & Mehrota, 2021). While literature examining therapists' mentalization is scarce, these findings may align with other research that has shown positive therapeutic effects of high therapist mentalization. For instance, Safran et al. (2014) found that high therapist RF (as measured by the RFS) predicted ratings pertaining to the extent that ruptures in the alliance were addressed and their degree of resolution.

Critiques and Future Directions

Many dimensions of the multidimensional model (Luyten et al., 2020) are well represented on the MentS, such as explicit/controlled mentalizing, mentalizing based on external cues, cognitive perspective taking, and mentalizing that focuses on both oneself and others. For instance, in item #2, "When I make conclusions about other people's personality traits, I carefully observe what they say and do" emphasis is placed on reflecting on others, not by

thinking about their *internal* states, but trying to decipher their *external* cues, such as their actions and what they speak about. This item, along with item #7, "When someone annoys me, I try to understand why I react in that way" are good examples of controlled mentalizing in which an individual consciously and deliberately puts effort into reflecting about mental states. Additionally, item #12, "I can make good predictions of other people's behavior when I know their beliefs and feelings" is an exemplar of cognitively based mentalizing, specifically, belief-desire reasoning or one's capability to predict an individual's actions through an understanding of their beliefs and desires. There are other items that seem to be related to one's capacity to understand others' perspectives yet may not be good indicators of this skill (e.g., "People tell me that I understand them and give them sound advice"). Scores on this item are contingent on another person's inferences about the respondent as well as the nature of the relationship between the respondent and the person making this judgement, both of which introduce bias into this assessment. Last, there are a few items that seem too generic to discriminate between those with low versus high mentalization capacities ("I often talk about emotions with people that I am close to", "I do not like to think about my problems.", etc.). Dimensions which appear to be excluded from the MentS are implicit/automatic mentalizing, mentalizing based on internal cues, and affectively based mentalization which involves automatic empathic processing and shared sensations of others' emotions. While there are items that focus on others' affect (e.g., item #6 "I can sympathize with other people's feelings") or seem to allude to emotional contagion (e.g., item #12 "Sometimes I can understand someone's feelings before s/he tells me anything"), neither captures the essence of the "implicit, visceral, bodily-based… system" based on "shared representations of others' mental states" (Luyten et al., 2020, p. 6). While item 12 seems to reflect a fast and possibly automatic process (a feature of the shared representation system) and

approximates the above definition, the phrasing "understanding someone's feelings" is not synonymous with *experiencing*, in one's body, another person's emotional experiences, and thus, falls short of capturing affective mentalizing as conceptualized by Luyten et al. (2020).

In conclusion, many findings support construct validity and internal reliability of the MentS. To expand the utility of this measure, it is recommended that future research investigates convergent validity with other measures of mentalization and further investigates the factor structure.

The Interactive Mentalizing Questionnaire

Development and Factor Structure

Following the development of the MentS, another team of researchers (Wu et al., 2022) sought to develop a measure that would capture mentalization-based processes involved in social interactions. They identified the following integral factors in these processes: 1) meta-cognition, that is, thinking about, perceiving, and being aware of one's own mental states, 2) perspective taking, defined as mentalizing about other's mental states, and 3) meta-mentalization, which is defined as the degree of insight individuals think others have about their ability to accurately represent mental states. The authors assert that while meta-cognition and perspective taking are represented in other measures, meta-mentalization has yet to be captured by other questionnaires. To develop the Interactive Mentalizing Questionnaire (IMQ), the authors generated 24 items intended to reflect the above three components. Four items were deleted for various reasons (e.g., low factor loadings and considerable conceptual overlap) for a total of 20 items. The authors performed EFA which specified a three or four factor model and consequently performed PCA, which indicated that three factors explained ~51% of the variance in responses. CFA also supported a three-factor model, with the following factors: meta-cognition, perspective taking

and meta-mentalization. Inter-factor correlations were strong. The perspective taking subscale were positively correlated with the meta-cognition subscale (r = .45) and negatively correlated with the meta-mentalization subscale (r = -.25). Factor loadings ranged from .53 – .81 with the exception of a lower loading (.38) for an item on the perspective taking subscale.

Reliability and Validity of the IMQ

The relatively moderate inter- factor correlations lend credence to the author's assertion that while these components all reflect interactive processes, they represent distinct elements of social interactions. Given that the meta-mentalization scale is reverse scored, the negative association between perspective taking and meta-mentalization signifies that individuals may overestimate both their ability to understand others' mental states as well as their ability to conceal their own mental states. The authors maintain that the positive relation between perspective taking and meta-cognition aligns well with stimulation theory (Harris, 1992), a framework which espouses the idea that individuals engage in a similar cognitive process when constructing others' mental states in their minds and making inferences about their own mental states. Lending evidence to the internal consistency of the subscales, Cronbach's alpha values ranged from .76 - .83. The authors also provide some evidence for construct validity; their results demonstrated negative correlations ($-.31 < r < -.42$) between the subscales on the IMQ and the subscales on the Autism Spectrum Quotient scale (ASQ; S. Baron-Cohen et al., 2001). This finding is consistent with theoretical and empirical investigations suggesting that individuals with Autism Spectrum Disorder experience deficits in meta-cognition and mentalization (Zalla et al., 2015). The meta-cognition subscale had the highest negative correlation with the ASQ, a finding that is corroborated by previous research suggesting that individuals with ASD have considerable difficulty identifying their own emotions and thoughts. Conversely, some results

failed to support construct validity. None of the subscales on the IMQ were correlated (-.05 < *r* <.15) with the Empathic Concern (EC) scale from the Interpersonal Reactivity Index (IRI; Davis, 1983), a well-established measure of empathy as a multidimensional construct. This is a surprising and concerning finding given the considerable conceptual overlap between empathy and mentalization in reference to others' internal states (Choi-Kain & Gunderson, 2008).

To further examine construct and ecological validity of the IMQ, the authors conducted a social experiment where participants were assigned to the role of either proposer or responder. The proposer received a small amount of money and was instructed to disclose the amount and to offer the responder some portion. The responder could accept or reject the offer. The authors found that proposers' meta-mentalization scores were higher after their offers were accepted than when they were rejected. Of the proposers, those who believed others did not have insight into their mental states (higher meta-mentalization) had higher confidence ratings and perceived the portioned rewards to be fairer than those with lower meta-mentalization scores. Regarding the responders, those with higher meta-mentalization did not trust their proposer, however those with higher perspective taking were more likely to trust the proposer. In view of these findings, Wu and colleagues' (2022) assert that the IMQ is sensitive to changes in social interactions and is associated with important perceptions and behaviors during social situations.

Critiques and Future Directions

Despite these promising findings, there are many limitations of the IMQ as a measure of mentalization in social interactions. First and foremost, the authors did not provide an explanation regarding the manner in which they generated the items on the IMQ, nor the source of their content in the first publication of this measure (Wu et al., 2022). A vital step in scale development is to consult with experts in the field and request that they review and judge the

(face) validity of the proposed test items. However, Wu and colleagues' (2022) did not refer to experts in the areas of mentalization and socio-communicative process in order to finalize their test items. Moreover, one could argue that the null findings between the IMQ and IRI may be due to the scales on the IMQ capturing more of the cognitive, as opposed to, affective elements of mentalization (the affective component being heavily reflected in the EC scale on the IRI). This remains an open question but seems like it could be an adequate explanation based on the face validity of scale items, which seem to capture cognitive (rather than affective) content and processes of mentalization (e.g., item 3, "I believe that I am good at telling what another person is thinking"). Indeed, there are no items on this measure which refer to understanding one's own or others' affect. While cognitively based mentalizing appears to be captured by such general items on the IMQ, there is no content which represents belief-desire reasoning or perspective taking. It is plausible that the lack of specificity could represent automatic mentalizing (e.g., having an intuitive sense of another person's internal world instead of making deliberate and effortful attempts to understand one's thoughts and feelings). In view of the nature of items which refer to mentalizing about others (e.g., item 3 and item 4, "I'm confident that I can tell what others are thinking"), emphasis is placed on internally based mentalizing rather than trying to decipher others' external cues.

As of the writing of this thesis (August 2022), the IMQ has not been widely adopted and the seminal paper by Wu et al. (2022) has only been cited three times. Regardless, this measure does seem to be applicable to dynamic interactive processes and examines a facet of mentalization (meta-mentalization) which has seldom been included in other measures. Future work on this measure should examine the association between the IMQ and other measures of mentalization.

The Multidimensional Mentalizing Questionnaire

Development and Factor Structure

Given that most self-report questionnaires do not adequately capture all the dimensions which constitute mentalization, Gori et al. (2021) sought to ameliorate this issue by creating a comprehensive multidimensional scale. The authors generated 33 items by referring to the Handbook of Mentalizing in Mental Health Practice (Bateman & Fonagy, 2012) and consulted with researchers and clinical experts in this area to develop the Multidimensional Mentalizing Questionnaire (MMQ). No additional details were provided about the process for developing the MMQ. In their initial work, Gori et al. (2021) recruited a large community sample (N = 349) in Italy and conducted EFA which indicated a six factor model that explained about 57% of the variance. The specified factors are *reflexivity*, which accounted for ~20% of the variance, *ego-strength* which explained ~17%, and the latter four factors *relational attunement, relational discomfort, distrust,* and *emotional dyscontrol*, which each explained between 4-6% of the variance. Reflexivity is described as one's tendency to explore, monitor, and understand mental states and an inclination towards deciphering the meaning of behaviors and events. Ego-strength refers to one's ability to tolerate and navigate everyday problems "with an emotional resistance to stress and frustrations" (Gori et al., 2021, p.12) and in turn, facilitates self-preservation and self-efficacy. The conceptual relation between ego strength and mentalization was discussed in Winkler's (2014) psychodynamically influenced paper where he asserted that the ego serves the function of structuring one's emotional experiences into a cohesive self-identity/personality and that one's capacity to mentalize is the process through which this function is served. Relational attunement is described as being in tune with others' internal states (including emotions and cognitions) and refers to a similar phenomenon, a shared representation (SR) system, that was

highlighted by Ripoll et al. (2013) and was later framed as affective mentalizing in Luyten et al.'s (2020) review paper. Specifically, they explain that in the SR system, an individual's observations of another's emotional experiences or mental states trigger mirror neurons to automatically mimic the neuronal activity that is linked to emotional/cognitive experiences, allowing shared sensations of others' emotions. Similarly, Gori et al. (2021) refer to Waal's (2006) description of subject-object state matching which bears a striking resemblance to the SR system; they assert that when an individual observes another person's emotional states, "neural representations of similar states are automatically activated" and specifically, there is an activation of similar "motor and autonomic responses…(e.g., changes in heart rate, skin conductance, facial expression, body posture)" (Waal, 2006, p. 37). They also highlight the role of mirror neurons in acting as a linking mechanism between perceiving others' states and one's own similar neural activation to the other. In sum, the relational attunement factor bears conceptual similarity to Luyten et al.'s (2020) description of affectively based mentalizing.

The first three factors of the MMQ, namely *reflexivity, ego-strength,* and *relational attunement*, are described as "good mentalizing" and refer to functionally adaptive facets of mentalizing. In contrast, the latter three factors, *relational discomfort*, *distrust*, and *emotional dyscontrol*, broadly refer to "bad mentalizing" and are conceptualized as failures to mentalize. Relational discomfort refers to frequent difficulties with interpersonal relationships, feeling misunderstood and hurt by others and consequently isolating oneself due to fear of abandonment. The distrust subscale is characterized by close-mindedness, mental rigidity, and a tendency to distrust others in interpersonal relationships. Individuals high on this factor tend to engage in cyclical patterns of behavior where maladaptive perceptions of the self as delicate and others as intimidating are reiterated. As such they avoid new experiences from which social learning may

occur, a common behavior noted in individuals with BPD (Fonagy et al., 2015). Last, the emotional dyscontrol subscale specifies difficulties with emotion regulation and impulsivity.

Reliability and Validity of the MMQ

Gori et al. (2021) conducted two studies to examine the psychometric properties of the MMQ; the first was comprised of a large community sample (N = 349) and the second study compared scores on the MMQ for a small community sample (n = 50) and a clinical (n = 46) sample. In the first study, Cronbach's α internal consistency estimates ranged from .72 - .89. The correlations between the MMQ and various other measures were largely indicative of construct validity. For instance, alexithymia was negatively correlated with the first three "positive" mentalization factors ($-.33 < r < -.49$) and positively correlated with the three "negative" mentalization subscales ($.43 < r < .54$). The same trend was seen with correlations between the MMQ subscales and a measure of impulsiveness. As anticipated, the association with the emotional dyscontrol subscale, characterized by impulsive behavior, was the strongest (r = .42). Gori et al. (2021) also investigated the relations between the MMQ and a measure of self-efficacy (which assesses an individual's beliefs about their ability to cope with difficult problems) and the opposite pattern of associations was observed (positive correlations with the first three "positive" mentalization factors and negative correlations with the last three subscales), although correlations were of lower strength ($|.20| < r < |.37|$) with the exception of ego-strength. In line with the strong conceptual similarities between self-efficacy and the ego-strength subscale, the strongest association emerged with this subscale (r = .69). Relations of the MMQ with self-esteem scale were more complex. From a theoretical standpoint, it is reasonable to assume that one's capacity for mentalization and one's self-esteem would be largely unrelated and Fonagy et al.'s (1998) findings corroborated this assumption. In contrast, Gori et al. (2021)

found significant negative relations with the three "negative" subscales, relational discomfort, distrust, and emotional dyscontrol (-.35 < *r* < -.56) and a positive relation with ego-strength (*r* = .48). Self-esteem was unrelated to reflexivity (*r* = .08) and relational attunement (*r* = .05). These results are difficult to interpret because this association has seldom been investigated, however the findings may suggest that the MMQ might be inappropriately capturing self-esteem.

The associations between the MMQ subscales and secure attachment were in the expected directions (positive associations with three "good" mentalizing factors (.22 < *r* < .35) and negative relations with relational discomfort (*r* = -.37) and discomfort (*r* = -.16)). However, secure attachment was unrelated to emotional dyscontrol (*r* = -.09). Associations between the MMQ and insecure attachment styles (preoccupied, avoidant, and unresolved) were also in anticipated directions (|.01|< *r* < | .51|). Most of the associations with avoidant attachment style did not reach significance (|.01|< *r* < | .1|). Last, the authors found that extraversion, agreeableness, openness, conscientiousness, and emotional stability were positively related to the first three "positive" mentalizing factors (.10< *r* < .51), and negatively related to the three "negative" mentalizing factors (-.06 < *r* < -.48), with a few exceptions. While these findings mostly align with Dimitrijević et al.'s (2018) work with personality variables and the MentS, they are in contrast to seminal work by Fonagy et al. (1998) who found that extraversion and neuroticism were unrelated to the RFS. Although some research has found associations between task-based measures of mentalizing and Big Five personality traits (Nettle & Liddle, 2008), little research has investigated this association with self-report measures of mentalization. It is possible that the various dimensions of mentalizing may have different associations with personality variables. For instance, statements indicative of automatic mentalizing were excluded from both the MentS and MMQ yet automatic mentalizing may be one of the dimensions of

mentalization that is elicited by the RFS, given the lack of constraints placed on their narratives. Perhaps automatic mentalizing is less related to personality traits than other dimensions of mentalization, which may explain why Fonagy et al. (1998) did not find a significant relationship but Gori et al. (2021) and Dimitrijević et al. (2018) did.

In their second study, Gori et al. (2021) examined differences in MMQ scores between a community sample and a clinical sample composed of individuals diagnosed with psychotic, mood, anxiety related, personality, and obsessive-compulsive disorders. They found that the clinical sample had significantly lower scores on the total MMQ and all subscales except relational attunement and emotional dyscontrol, partially supporting construct validity. The null findings regarding the above-mentioned factors may suggest these sub-capacities of mentalization were not extremely impaired in this clinical sample or may be explained by the small sample sizes. More encouragingly, the significantly lower scores on the other subscales for the clinical group is consistent with theoretical assumptions. Indeed, previous research has demonstrated that individuals with schizophrenia, personality disorders, and bipolar disorders have impaired metacognitive abilities and difficulties in understanding their own and others' feelings and intentions as well as having extreme conviction in their beliefs about themselves as well as others. Overall, initial validation studies support construct validity and internal reliability of the MMQ, however, this measure has not yet been used in any other empirical work to date.

Critiques and Future Directions

A common weakness of the self-report measures presented thus far is that not all of the dimensions of mentalization (e.g., implicit vs. explicit, self vs. other, internal vs. external, and cognitive vs. affective) are sufficiently captured. Gori et al. (2021) sought to rectify this issue; their visual model asserted that the three "good" mentalizing factors are characterized by

"cognitive-affective integration" and "internal-external openness", with implicit and explicit mentalizing on a continuum ranging from reflexivity (implicit) to ego-strength to relational attunement (explicit). Despite the appeal of a multidimensional model of mentalizing, there are various issues with the framework presented by Gori et al. (2021). First, it is debatable whether cognitive-affective integration is represented across the items in each of the three factors. For instance, most of the items in the reflexivity factor identify cognitive features of mentalization, (e.g., "Understanding what others feel is crucial in understanding their actions") yet do not include "embodied affective features that ground mentalizing in an affectively felt reality" (Luyten et al., 2020, p. 6). A parallel issue exists with the relational attunement subscale; for instance, perspective taking is represented yet affective mentalizing is not. As mentioned above, the relational attunement subscale is similar to Luyten et al.'s (2020) conceptualization of affective mentalizing, yet none of the items adequately capture the essence of the shared representation system where an individual experiences another person's emotions through the mirroring of internal states. Moreover, Gori et al. (2021) placed relational attunement towards the explicit end of the implicit/explicit continuum, which implies that this is more of a controlled and reflective process, however, Luyten et al. (2020) specify that affective mentalizing is largely a fast and automatic process that relies on older brain circuits. Since relational attunement is one of the three "good" mentalizing factors, it is supposedly characterized by "internal-external openness" which refers to mentalizing that is grounded in examining both internal and external features. However, it appears that all items in the relational attunement subscale are focused on interpreting others' internal states through perspective taking (e.g., "I can tune in other people's mental states") which imply mentalizing from an implicit stance. Attempts to understand others through scrutinizing external factors, such as facial expressions, are not represented. Thus, the

authors' claim that the relational attunement scale demonstrates "internal-external openness", "cognitive-affective integration", and is explicitly oriented, is dubious. Regarding the three "bad" mentalizing factors (relational discomfort, distrust, and emotional dyscontrol), Gori et al. (2021) asserted that these factors were characterized by "internal-external closure" and "cognitive-affective split". These latter three subscales are unique in that they do not contain items that directly assess one's ability to understand their own as well as others' mental states but instead are a reflection of *consequences* of lower mentalizing skills or mentalizing that is over-reliant on certain dimensions over others. This assertion is in line with Luyten and colleagues' (2020) theoretical tenets stating that imbalances along mentalization dimensions refers to excessive functioning of one polarity and a corresponding dysfunction in the opposite polarity. However, Gori et al. (2021) do not specify how the various dimensions are imbalanced (i.e., which pole is dominant and which is functioning to a weaker degree), with the exception of the implicit/explicit (or automatic/controlled) polarity. It could be argued that items on the relational discomfort subscale (e.g., "Others don't understand me") align with judgements about others' mental states and intentions that are predominantly implicit or automatic in nature. Other potential weaknesses of the MMQ are 1) the distrust and emotional dyscontrol scales are composed of only four items each and thus, may not capture these underlying constructs to a satisfactory degree, and 2) there are a few items which appear too general to discriminate between individuals of low vs. high mentalizing capacities (e.g., "I am a thoughtful person" and "I ponder over what happens to me"). Overall, it is laudable that Gori et al. (2021) aimed to represent the four dimensions of mentalizing yet various issues with their conceptual model have been highlighted. Further validation studies with the English translation of this measure and larger clinical samples are warranted before the MMQ is widely used across this literature.

Current Study

Through examining both the theoretical underpinnings of mentalization and its operationalization it is evident that there is an incongruence between the great strides in the conceptual development of this construct and the comparatively minimal empirical work surrounding its construct validity. The paucity of research centered on the convergence of mentalization measures is problematic because researchers use these tools interchangeably with the assumption that all measures assess the same underlying construct. It seems possible, however, that these tools may reflect related but not identical constructs. For instance, the conceptual boundaries of this construct are not well-defined and researchers may deem different combinations of psychological capacities (e.g., empathy, perspective taking, emotion recognition and regulation) as being involved in mentalizing. As a result, there are likely differences in which constructs/components are represented across measures, depending on preferences of different author teams. Similarly, research teams have drawn upon distinct yet related conceptual frameworks to develop their measures. In particular, Gori et al. (2021) referred to the multidimensional model of mentalizing (Luyten et al., 2020) to generate the item content and structure of the MMQ whereas Hausberg et al. (2012) referred to the theory related to pre-mentalizing modes (Fonagy and Target, 1996), which is largely based on clinical anecdotes from working with BPD clients. Last, Fonagy et al. (2016) drew on dual deficit model of mentalization (e.g., hypomentalization and hypermentalization) to organize the RFQ. The distinct yet related frameworks on which the self-report measures are based further highlight the need for empirical investigations into convergent validity.

In addition to the limitations of the self-report measures discussed thus far, another prominent shortcoming is that one must have a strong capacity for mentalizing in order to

accurately report on their tendencies to mentalize across various contexts (Hausberg et al., 2012). As such, individuals who are predicted to have severe deficits in mentalizing (e.g., clients with personality disorders) may not have insight into their difficulties with mentalizing and thus, may introduce bias into their reporting (Gagliardini et al., 2018). In view of this important drawback, Luyten et al. (2012) have endorsed various experimental/observational tasks to assess mentalization, such as the Movie for the Assessment of Social Cognition (MASC; Dziobek et al., 2006). These types of measures are useful because they circumvent the issue of individuals' inferring their own mentalization skills and use standardized procedures which require minimal training to administer (Fossati et al., 2018). The MASC has been found to have good internal reliability (alpha > .80), test-retest reliability (r = .97) and construct validity (e.g., individuals with Asperberger Syndrome and BPD scored significantly lower than non-clinical samples) (Dziobek et al., 2006; Preissler et al., 2010). Given that the MASC has been endorsed by the same set of authors (Luyten and Fonagy) who pioneered theoretical and empirical research on mentalization, one of the aims of the current study was to investigate the relations between common self-report measures and the MASC.

To the authors' knowledge, only 1 study has examined the associations between the self-report measures of interest in the current study. Raimondi et al. (2021) found that the Italian translated MZQ was strongly associated with the RFQ_C (r =.60, p<.001) and negatively correlated (r = -.41, p<.001) with the RFQ_U. Furthermore, only four studies have investigated the associations between the MASC and the self-report measures used in the current study (Duval et al., 2018; Schwarzer et al., 2021a; Sharp et al., 2021; Rothschild-Yakar et al., 2019). Both Duval et al. (2018) and Sharp et al. (2021) investigated the relations between the MASC and the Reflective Functioning Scale for Youth (RFQ-Y; Sharp et al., 2009) which is an

adaptation of the adult version of the RFQ. The RFQ-Y contains two subscales which assess adolescents' awareness and understanding of their own and others' emotions, thoughts, and behaviors. Sharp et al. (2021) found a non-significant correlation (r =.13) between the total score on the MASC and one of the subscales on the RFQ-Y. Duval et al. (2018) conducted an exploratory factor analysis on the RFQ-Y and the results were consistent with a 3-factor structure which included 1) uncertainty about mental states, 2) interest/curiosity about mental processes and 3) excessive certainty about mental states. When analyzing the associations between these factors and the MASC, the authors found a mix of significant and non-significant results. The uncertainty subscale of the RFQ-Y was not significantly associated with any of the MASC scales whereas the interest/curiosity subscale was significantly negatively associated ($-.31 < r < -.23$) with all types of errors on the MASC (e.g., hypermentalizing, hypomentalizing, and no mentalizing) (see Measures). Excessive certainty was marginally, although significantly associated with hypomentalizing errors (r = -.18) on the MASC but no other scales. The results of Rothschild-Yakar et al.'s (2019) analyses were also mixed. They found similar non-significant results ($-.24 < r < .24$) regarding the association between the RFQ_C and all types of errors on the MASC yet the RFQ_U was significantly associated with hypermentalizing errors (r = .38, p<.01) and the total score (r = -.31, p <.05) on the MASC. Schwarzer et al. (2021a) found a small yet significant correlation (r = .24, p <.001) between the MZQ total score and number of correct answers on the MASC. Taken together, these studies may suggest a lack of convergence between self-report and task-based measures of mentalizing. While current evidence is inconclusive, it is worth highlighting that researchers have reported weak associations between scores on self-report measures and behavioral/ experimental tasks assessing cognitive empathy (Murphy &

Lilienfeld, 2019), a core component of mentalizing and recommend that self-report questionnaires should not be used as proxies for task-based measures.

In view of the ill-defined bounds of this construct and the scant empirical evidence regarding convergent validity, the current study has two primary aims: 1) to examine the associations between self-report tools and a task-based measure of mentalization, and 2) to conduct an exploratory factor analysis to identify the common latent factors underlying all five self-report measures at the subscale level. In view of previous research examining the convergence of various measures of mentalizing, the following hypotheses will be tested.

1. *All self-report measures will be highly correlated, with some exceptions.*
    a. Based on Raimondi et al.'s (2021) findings, *the MZQ will be moderately positively correlated with the RFQ_C and moderately negatively correlated with the RFQ_U.*
2. *All self-report measures will be moderately associated with performance on the MASC,* given that prior research has found low convergence between most self-report measures and behavioral tasks assessing cognitive empathy (Murphy & Lilienfeld, 2019).
    a. Based on Schwarzer et al.'s findings (2021a), *the MZQ will be positively correlated with the MASC.*
    b. *The RFQ_C will be highly positively correlated with and will significantly predict hypermentalization errors on the MASC. The RFQ_U will be strongly correlated with and will significantly predict hypomentalization errors on the MASC.*

Both of the overarching hypotheses are drawn from the fact that all self-report tools and the MASC are situated in the same theoretical framework pertaining to mentalization. While findings are mixed regarding the association between the RFQ and the MASC, hypothesis 2b follows from the fact that the RFQ and the MASC both map onto the theoretical framework which stipulates that there are two types of mentalization deficits: hypomentalization and hypermentalization. Given the paucity of work examining the convergent validity of the self-

report and task based measures, explicit hypotheses regarding the second aim of this project were not put forward as these analyses were exploratory in nature.

Methods

Study Sample

In the current study, participants were undergraduate psychology students in the University of Waterloo Research Experiences Group (REG) participant pool recruited through the SONA system. There were no inclusion or exclusion criteria with the exception that participants had to be able to watch and listen to videos during their participation. A total of 348 participants completed the study. Data integrity analyses led to a reduction of the sample to 247 participants (77.64% female with a mean age of 21.62 [SD = 3.13]). Table 1 presents further demographic information.

Procedures

The study was reviewed and approved by the research ethics board at the University of Waterloo. Psychology students read a brief description of the study on SONA and signed up if they were interested. Participants then clicked on a Qualtrics link which directed them to the survey, which included the consent form, three self-report measures of mentalizing, and a task-based measure (MASC; Dziobek et al., 2006). All participants completed MZQ (Hausberg et al., 2012), the RFQ (Fonagy et al., 2016), and the MASC (Dziobek et al., 2006). Last, they completed a third self-report measure, which was randomly chosen from three possibilities: the MentS (Dimitrijev et al., 2018), the MMQ (Gori et al., 2021), or the IMQ (Wu et al., 2022). After completing all study measures, they were presented with the feedback letter. Participants received 1.0 SONA credit for completing the study. Participants were only asked to complete one of the three latter measures to ensure that the total time required for the study was within reasonable limits (e.g., less than an hour) to maximize participants' attention and motivation to complete the study. The RFQ and MZQ were chosen as the two measures that all participants

completed because they are the most widely cited self-report measures and were generated based on core theoretical tenets (e.g., pre-mentalizing modes for the MZQ and the dual deficit model for the RFQ).

To ensure acceptable data quality, various procedures were employed to trim the dataset and remove data from participants who completed measures too hastily and who provided multiple identical responses in a subsequent fashion. In total, 101 cases were deleted from the dataset. Of these 101 cases, 80 were deleted because they completed the MASC (Dziobek et al., 2006) in 900 seconds or fewer. Given that the total length of the videos was 946 seconds, the authors established the criterion of 900 seconds as it included participants who may have submitted their answers a few seconds before the end of longer video clips but removed participants who bypassed a few videos at minimum. Subsequently, 21 cases were removed due to rapid completion of one or more self-report measures (RFQ, MZQ, IMQ, MentS, or MMQ). These cases were deleted from the dataset based on the guideline advocated by Huang et al. (2012) who suggested that participants' responses are likely of poor quality if they spend less than 2 seconds per question. In the last procedure, the authors conducted a longstring analysis where a variable was created which represented the number of consecutive identical responses for each measure. For this criterion, data for the particular measure in which there was a consistent response style, was deleted. Entire cases were not deleted if participants appeared to provide data of sufficient quality for other measures. Specifically, RFQ data was removed for three participants who had 6 consecutive identical responses (out of a total of 8 questions). MZQ data was deleted for two participants who had 9 and 10 identical subsequent responses (out of 15 questions) and MentS data was removed for one participant who had identical responses for 14 questions (out of 28 questions). To determine the number of identical responses which warranted

removal, the authors examined histograms and frequency tables for the longstring variables associated with each measure. All distributions of the longstring variables were positively skewed and outliers were removed. These filters resulted in a final sample of 247 participants. Given that the participants were randomly assigned to complete one of three final self-report measures (the MentS, MMQ, or the IMQ), analyses which include any of these three measures have markedly smaller sample sizes than analyses which exclusively include the RFQ, MZQ, and/or MASC.

Measures

Self-report Measures

While the response scales differed across the measures as described in the original publications, the authors implemented a universal response key (specified below) for all self-report measurement tools in order to prevent groups of items and overall measures from aggregating together in the exploratory factor analyses simply because of similar response keys. Moreover, while the response options for all measures was indicative of level of agreement with items, the labels and the direction of the scale differed across some measures (e.g., the IMQ ranged from (1) *Very true for me* to (5) *Very false for me* while the MentS ranged from (1) *Completely Incorrect* to (5) *Completely Correct*). In the current study, the researchers implemented the following Likert scale ranging from (1) *Strongly Disagree* to (5) *Strongly Agree* for all self-report tools. A 5-point scale was selected because the majority of the measures were originally rated on a 5-point scale (i.e., the MZQ, MentS, and MMQ).

Participants first completed the *Reflective Functioning Questionnaire* (RFQ; Fonagy et al., 2016), an 8-item instrument which assesses the degree to which individuals believe they understand and can identify their emotions, motivations for their actions, and their ability to

regulate their emotions. The content of all items except for #7 implies difficulties with mentalizing (e.g., "I don't always know why I do what I do"). In accordance with the instructions for the measure, the RFQ was scored in two ways. In the first method, two subscales were calculated: Certainty about Mental States (RFQ_C) and Uncertainty about Mental States (RFQ_U). The RFQ_C measures degree of overconfidence in assessing one's own and other's mental states and includes items 1-6. For the RFQ_C, the scale was re-scored from (1) *Strongly Disagree* to (5) *Strongly Agree* to 3, 2, 1, 0, 0 where strong disagreement with statements is indicative of higher levels of certainty. The RFQ_U assesses the extent of difficulties in understanding and inferring mental states and includes the following items: 2, 4, 5, 6, 7, 8. Apart from item 7, items are re-scored from (1) *Strongly Disagree* to (5) *Strongly Agree* to 0, 0, 1, 2, 3 where strong agreement with statements is indicative of higher levels of uncertainty. Given that item 7 is phrased to indicate hypermentalizing, it was scored according to the scale for the RFQ_C (e.g., 3, 2, 1, 0, 0). It's noteworthy that items 2-6 are included in each subscale, resulting in redundant information. These scoring procedures were recommended by Fonagy et al. (2016) who did not provide their rationale for collapsing the response scale. The second method of scoring simply involved reverse coding the original response scale from 1-5 for all items except 7 so that lower responses indicated uncertainty and higher responses indicated certainty (Müller et al., 2021). A total score was calculated through summing all items from this latter scoring method and it is referred to as the RFQ sum throughout analyses.

The *Mentalization Questionnaire* (MZQ; Hausberg et al., 2012) was the second measure presented to participants. It is a 15-item measure which includes four subscales, 1) refusing self-reflection, defined as avoiding thinking about mental states and being fearful of being overwhelmed by affective experiences, 2) emotional awareness, in which items are indicative of

deficits in identifying and differentiating between mental states, 3) psychic equivalence where one's thoughts and feelings are experienced as real events that are bound to occur instead of mental representations and hypotheses about what may happen, and 4) regulation of affect where items represent difficulties controlling one's emotional experiences. All items were reverse scored and as such, higher scores represented lower levels of difficulties in these domains and adaptive mentalizing. Items within each subscale were summed to calculate four scores. In addition to the subscale scores, item scores were summed together to create a total score with higher scores indicative of a better capacity for mentalizing.

The *Mentalization Scale* (MentS; Dimitrijev et al., 2018) is a 28-item questionnaire which includes three subscales: 1) other-related mentalization (MentS-O), which refers to understanding and making inferences about others' mental states, 2) self-related mentalization (MentS-S), assessing one's insights into their own beliefs, desires, motivations, etc., and 3) motivation to mentalize (MentS-M), defined as one's interest in understanding mental states and reasons for individuals' behaviors. Three subscale scores were constructed by summing items in each subscale and the total score was generated by summing all the items. Higher values on both the subscale and total scores are indicative of better mentalizing.

The *Multidimensional Mentalizing Questionnaire* (MMQ; Gori et al., 2021) is a 33-item measure intended to capture various factors situated in an integrated and multilevel model of mentalizing, which includes the following dimensions: self-other, internal-external, cognitive-affective, and implicit-explicit. There are six subscales included in this measure, three which assess adaptive facets of mentalizing, namely reflexivity, ego-strength, and relational attunement, and three which measure deficits in mentalizing, specifically, relational discomfort, distrust, and emotional dyscontrol. Subscale scores were constructed through summing applicable items and

higher values on the first three subscales are indicative of better mentalizing. In contrast, higher scores on the latter three subscales are reflective of poor mentalizing. To generate a global score, the relational discomfort, distrust, and emotional dyscontrol subscales were reverse scored and summed with the other three subscales.

The *Interactive Mentalizing Questionnaire* (IMQ; Wu et al., 2022) is comprised of 20 items and includes three subscales: meta-cognition, perspective taking, and meta-mentalization. Meta-cognition refers to assessing and understanding one's own thoughts, desires, etc. Perspective taking is defined as interpreting and understanding others' mental states. Meta-mentalization refers to the degree to which an individual believes that others can infer one's own thoughts, desires, and intentions. Items contained within each subscale were summed to calculate three scores and a total score was calculating by summing all items. For subscale and total scores, higher values are reflective of a high capacity for mentalization.

Sum scores for all self-report measures were calculated only if ~60% of the items that made up the sum, were completed. In the case of missing responses, sum scores were pro-rated based on the number of completed items. To generate pro-rated sum scores, the total sum was divided by the number of completed items and subsequently, multiplied by the total number of items on the measure.

Task-based Measures

Participants completed the *Movie for the Assessment of Social Cognition* (MASC; Dziobek et al., 2006) after the RFQ and the MZQ. The MASC includes 51 multiple choice items which are based on short video clips. The total running time of clips is approximately 15 minutes and the videos are centered around four adults having a dinner party. There are 43 individual clips and 1-2 multiple choice questions after each clip. After being introduced to the characters,

participants are asked to answer questions about the characters' thoughts, feelings, and intentions (e.g., “What is Sandra feeling?”). The videos are originally in German but have been translated into English and the English translation has been used across various research studies (Hezel & McNally, 2014; Montag et al., 2011; Sharp et al., 2011). There are four response options for each multiple choice question, including three types of error and one correct answer. The three categories of error are 1) hypermentalizing, referring to an overinterpretation of mental states (such as one’s motivations for making a specific statement) 2) hypomentalizing which is indicative of overly simple inferences about the characters’ mental states, and 3) no mentalizing, which reflects fact or situation based inferences to explain characters’ behaviors and makes no reference to their mental states.

Analytical Approach

To address the research questions about the associations between the self-report tools and the MASC and to conduct various exploratory factor analyses on these measures, the Statistical Package for the Social Sciences (SPSS) version 28 was used.

Results

Associations Between Self-Report Measures and the MASC

To address the first aim of this study, bivariate correlations were calculated between all self-report measures and between these questionnaires and the three types of errors as well as correct responses on the MASC (see Tables 2 -7). Alpha coefficients are also provided in these tables. Hypothesis 1 was partially supported. In view of the high internal reliability of most scores (e.g., alpha > .70), correlations above .60 are indicative of strong associations between the measures. As such, strong relations emerged between various sum level scores (see Table 2), including the association of the MMQ with both the RFQ (r = .65) and MZQ (r = .66) and the relation between the MZQ and RFQ (r = .62). These strong associations suggest that the RFQ, MZQ, and MMQ are measuring facets of a common latent variable. In contrast, the IMQ appears to be moderately related to the RFQ (r = .47) and weakly linked to the MZQ (r = .33). These findings suggest that the IMQ and MZQ are not measuring the same latent construct however, the IMQ and the RFQ are likely capturing some similar components. The MentS appears to be moderately associated with the RFQ (r = .46) and the MZQ (r = .51), suggesting that these measures are tapping some degree of overlapping content, but it is debatable whether they are measuring the same latent construct. Regarding hypothesis 1a, findings replicated Raimondi et al.'s (2021) results pertaining to the association between the MZQ and RFQ_C (r =.54) yet the relation between the MZQ and RFQ_U was stronger in the current project (r = -.60) than this prior work (Table 2). Of note, the RFQ_C and RFQ_U are very highly negatively correlated (e.g., bivariate correlation is -.86 and the disattenuated correlation is -1.16 which is clearly an overestimate of -1), suggesting that these subscales are highly redundant and do not warrant

separate measures in these analyses. As such, the sum-level score of the RFQ was used for the remainder of analyses.

Most of these trends at the sum level were consistent at the subscale level, with lower internal reliability estimates and correlations overall. Correlations ranged from .03 - .65 with the majority of them lower than .4. Most correlations between the RFQ and the MMQ and between the RFQ and MZQ ranged from .40 - .60 (see Table 3), suggesting strong associations given the lower internal reliability values which restrict the correlations. In contrast to the sum level findings, most correlations between the MMQ and the MZQ subscales were under .25 with one notable exception, the association between the Affect Regulation subscale on the MZQ and the Emotional Dyscontrol subscale on the MMQ (bivariate correlation is r = -.73 and the disattenuated correlation is -1.31, again an overestimate of -1). Similar to the moderate associations noted above regarding the IMQ and RFQ, associations between subscales of these measures (see Table 4) were all below .34. A similar pattern was observed with the IMQ and MZQ where all correlations were below .30 (see Table 5). Last, an interesting trend emerged regarding the associations between the MentS with the RFQ (see Table 4) and the MZQ (see Table 6). The relations between the MentS_O and the MentS_M with all subscales on the RFQ and MZQ were below .23 however associations with the MentS_S were above .45. These findings may imply similarities (e.g., in the construct being captured and/or themes of items) between the RFQ, MZQ and the MentS_S, but not with the MentS_M and MentS_O. Of note, the associations between the IMQ, MMQ, and MentS could not be calculated given that participants only received one of these measures.

Hypothesis 2 regarding the association between self-report measures and the MASC was not supported. All of the self-report measures had very weak linkages with the MASC (see

Tables 2, 4, 6, and 7). While a few of the correlations between the self-report measures with the four types of responses on the MASC were statistically significant, they were low/weak ($r < .29$). Additionally, hypothesis 2a predicting a similar strength correlation ($r = .24$) as Schwarzer et al. (2021a) regarding MASC performance and the MZQ was partially supported ($r = .111$, $p = .085$, 95% CI [-.015, .234]). Last, hypothesis 2b regarding the association between RFQ subscales and types of MASC errors was not supported. The RFQ_C was not associated with hypermentalization errors on the MASC ($r = -.051$, $p=.427$) nor was the RFQ_U related to hypomentalization errors ($r = .065$, $p = .311$).

Self-Report Measures as Predictors of Performance on the MASC

Twenty linear regressions were estimated to investigate whether higher scores on self-report measures of mentalization predict better performance on the MASC (see Tables 8 and 10). In the first four models, predictors were the sum scores of self-report measures and the subsequent sixteen models included subscale level predictors. It was not possible to estimate a regression model with all self-report measures because in addition to the RFQ and MZQ, participants only completed one of three of the following measures: IMQ, MMQ, or MentS. The first regression model with sum scores of the RFQ and the MZQ as sole predictors was not significant ($R^2 = .014$, $F(2, 236) = 1.66$, $p = .193$) we report the standardized regression coefficients for this model to aid comparison with subsequent models that add predictors. In the latter three regression models the IMQ, MentS, or MMQ was added as the third predictor in addition to the RFQ and MZQ. When the MMQ was added, the model remained non-significant ($R^2 = .021$, $F(3, 76) = .549$, $p = .650$). In contrast, the model reached significance when the IMQ was added ($R^2 = .096$, $F(3, 78) = 2.75$, $p = .048$) yet none of the direct effects of the RFQ, MZQ, and the IMQ significantly predicted MASC performance ($p< .05$). In the final model with the

sum-level scores, the MentS was added as a predictor and the model reached significance (R^2 = .128, F(3, 70) = 3.43, p = .021). The RFQ (standardized β = -.344, p = .016) and MZQ (standardized β = .422, p = .005) independently predicted performance on MASC, and the direct effect of the MentS was not significant. As such, it appears the MentS acts a suppressor variable in the prediction of performance on the MASC with the RFQ and MZQ as predictors. Specifically, classical suppressor variables are defined as predictor variables with zero or weak correlations with the outcome variables, moderate correlations with other predictor variables, and increases in the R^2 (or the explained variance in the outcome variable) when the suppressor is present in the model versus absent (Horst, 1941). The MentS is uncorrelated with MASC performance (r = -.047, p= .69) but moderately correlated with the RFQ (r = .458 p< .001) and the MZQ (r = .510, p<.001). Last, there was a noteworthy increase in R^2 when the MentS was added to the model (e.g., from .014 to .128) and the direct effects of the MZQ and RFQ emerged as significant in the model with the MentS. Of note, the standardized Beta coefficient of the direct effect of the RFQ switched signs (e.g., went from positive to negative) when the MentS was added to the model (see Table 8 Model 4). To further investigate this suppression effect, partial correlations between the RFQ and MASC performance and likewise, between the MZQ and MASC performance were calculated, while controlling for the MentS. Lending further support for the suppression effect of the MentS, the partial correlations with both the RFQ and the MZQ were stronger than corresponding zero-order correlations (see Table 9).

In the first model with subscale level predictors, the RFQ sum and the four MZQ subscales were entered as predictors (see Table 10). In the second model, all three subscales from the IMQ were added as predictors, totaling 8 predictors. In the third model, all six subscales from the MMQ were added as predictors, resulting in 11 predictors. In the fourth model, all three

subscales from the MentS were included as predictors in addition to the RFQ and MZQ subscales, for a total of 8 predictors. None of these models were able to explain variance in MASC performance (based on correctly inferring mental states). Of the remaining twelve subscale level models, one subscale from the IMQ, MMQ, or MentS was added to the model in addition to the RFQ and MZQ subscales. As such, there were 6 predictors in each of these twelve latter models.

Of the sixteen subscale level models, all models failed to reach significance ($p < .05$) with five notable exceptions. When the meta-cognition subscale of the IMQ, or self-related mentalizing was entered as a predictor in addition to the RFQ sum score and MZQ subscales, this model reached significance ($R^2 = .157$, $F(6, 75) = 2.33$, $p = .04$). Similarly, the model including the meta-mentalization subscale of the IMQ significantly predicted MASC performance ($R^2 = .172$, $F(6, 75) = 2.61$, $p = .024$). Across both models, none of the direct effects emerged as significant predictors of MASC performance. In model 14, the MentS_S was added as a predictor in addition to the RFQ sum score and MZQ subscales and this model reached significance ($R^2 = .191$, $F(6, 67) = 2.64$, $p = .023$). The RFQ sum score (standardized $\beta = -.361$, $p = .016$) and the Affect Regulation subscale on the MZQ (standardized $\beta = .300$, $p = .040$) emerged as significant independent predictors. Similarly, when the MentS_O subscale was added as a predictor to the basic model (including the sum score of the RFQ and MZQ subscales), the model was significant ($R^2 = .187$, $F(6, 67) = 2.53$, $p = .029$). The sum score of the RFQ emerged as the only significant direct effect in this model (standardized $\beta = -.321$, $p = .031$). Last, model 16 which included the MentS_M, RFQ sum score, and MZQ subscales as predictors emerged as significant ($R^2 = .192$, $F(6, 67) = 2.66$, $p = .022$). The RFQ sum score (standardized $\beta = -.308$, $p = .038$) and the Refusing Self-Reflection subscale on the MZQ (standardized $\beta = .281$, $p = .042$)

emerged as significant direct effects (see Models 14 – 16, Table 10). Similar to the sum-level regressions discussed above, the MentS scales acted as classical suppressor variables in predicting the MASC from a subscale level (Horst, 1941). In accordance with the definition of classic suppression, there was a noteworthy increase in R^2 when each of the MentS subscales were added to models 14 – 16 and some of the direct effects went from being non-significant (e.g., in Model 1) to significant. Similar to the sum level regressions, the Beta coefficients for the direct effects of the RFQ sum score switched signs (e.g., went from positive to negative) after the MentS subscales were added (compare model 1 to 14, 15, and 16). Providing further evidence that MentS subscales are indeed acting as suppressors, partial correlations between MASC performance and the RFQ sum score as well as the Refusing Self-Reflection and Affect Regulation subscale on the MZQ were stronger than corresponding zero-order correlations (see Table 11).

Exploring the Factor Structure of Self-Report Measures

To address the third research question, exploratory factor analyses (EFA) with principal axis factoring was conducted. To assess whether the data was suitable for EFA, the Kaiser–Meyer–Olkin (KMO) and Bartlett's sphericity tests were conducted to assess the degree of common variance among measures. According to Kaiser & Rice (1974), values above .60 on the KMO indicated suitability for factor analysis. Bartlett's (1951) test of sphericity determines whether the correlation matrix between variables of interest is significantly different from an identity matrix. The Kaiser–Meyer–Olkin (KMO) measure was .832 and Bartlett's test of sphericity was significant, indicating that the correlation matrix was not identical (i.e., $\chi 2 = 511.393$, $df = 21$, $p<.001$). These findings suggest that this dataset was suitable for factor analysis.

Prior to the EFA at the subscale level, eight testlet variables were calculated. Each testlet variable represented the sum of 3-4 items within a subscale. RFQ Certainty testlet 1 included items 1 and 3 and RFQ Certainty testlet 2 included items 2, 4, 5, and 6. Items were scored according to procedures outlined for the RFQ_C subscale (e.g., 3, 2, 1, 0, 0). RFQ Uncertainty testlet 1 included items 7 and 8 and RFQ Uncertainty testlet 2 included items 2, 4, 5, and 6. Items were scored based on RFQ_U scoring protocols (e.g., 0, 0, 1, 2, 3). Four testlet variables were created for the MZQ based on the four subscales.

In the first EFA, the eight testlet variables representing the RFQ and MZQ were entered. These instruments were chosen to extract the factor structure because they are the most widely utilized mentalization self-report tools and were constructed based on central theoretical underpinnings, namely pre-mentalizing modes (introduced by Fonagy and Target, 1996) in the case of the MZQ and the dual deficit model in the case of the RFQ. As such, it is reasonable to assume that these instruments contain core features of the construct of mentalization. A 2-factor structure emerged, explaining ~64% of the variance. In accordance with the notion that the RFQ_C and RFQ_U reflect redundant information (e.g., bivariate correlation is -.86 and the disattenuated correlation is -1.16), factor loadings for the RFQ Certainty testlet 2 (.87 on factor 1 and -.44 on factor 2) and RFQ Uncertainty testlet 2 (-.87 on factor 1 and .42 on factor 2) were nearly identical but had opposite signs. Given that these testlets include the same items with different scoring protocols, the identical strength but opposing sign of the factor loadings may suggest that a second factor was been extracted simply due to the high negative correlation of these testlets. The authors ran a second EFA with Varimax rotation and Kaiser normalization. A 2-factor structure was extracted and the eigenvalues were 2.27 for factor 1 and 2.12 for factor 2. This factor model explained ~55% of the variance. A similar trend was noted was noted

regarding the RFQ testlets (RFQ Certainty testlet 2 (.93 on factor 1 and .28 on factor 2) and RFQ Uncertainty testlet 2 (- .92 on factor 1 and - .29 on factor 2) (see Table 12). To further support the notion that a second factor was extracted due to the strong negative relationship between Certainty and Uncertainty testlet 2 and that these testlets reflect overlapping information, the zero-order correlation between these testlets was very strong (r = -.942, p < .001). As a result, the RFQ Uncertainty testlet 2 was dropped from the remaining EFA. After conducting a third EFA with the remaining 7 testlet variables, a 1-factor structure emerged, which explained ~49% of the variance with an eigenvalue of 3.44. The factor loadings ranged from .60 – .71 (see Table 13) with the exception of the Refusing Self Reflection subscale on the MZQ which had a loading of .48.

To determine the factor loadings of the other self-report measures (e.g., IMQ, MMQ, and the MentS) onto this extracted single factor, zero-order correlations were calculated between the factor score estimated from the analysis of the RFQ and MZQ and the twelve subscales of the remaining three self-report tools as well as the sum scores of these measures. While ten of these fifteen correlations were significant (|.24 | < r < |.71 |, p <.05), only five were of a similar strength to the factor loadings of the MZQ and RFQ testlets onto this factor. Specifically, the MentS_S, Ego Strength, Relational Discomfort, and Emotional Dyscontrol subscales as well as the sum score on the MMQ, had factor loadings above .57 (see Table 13).

In addition to the factor analyses conducted, the authors sought to conduct EFA at the item level. However, given that this analysis involved a large number of variables (i.e., a total of 23 items across the MZQ and RFQ), there was insufficient data (N =238) to extract a stable factor solution. As such, the authors aim to conduct these analyses in a later project after more data has been collected.

Discussion

Given that limited work has examined the concordance between measures of mentalizing and that this construct lacks well-defined conceptual boundaries, the current study sought to elucidate the relations between various self-report questionnaires and a task-based assessment. A secondary aim was to investigate common latent factors underlying common and newly developed self-report measures. Partially supporting the authors' predictions, a pattern of strong associations emerged among between some self-report measures however, none of the self-report measures were strongly associated with the task-based measure. Last, given that mentalizing is considered a multifaceted construct, it was surprising that a 1-factor structure best explained the internal structure of two prominent measures. These results have significant implications for empirical research within psychology where mentalization is a construct of interest and for prominent theorists involved in the proliferation of mentalization theory.

Another Look at Defining Mentalization

Using Convergent Validity of Measures to Elucidate Conceptual Boundaries

Among the strong associations between self-report measures, some of these relations may be explained by overlapping item content, which may suggest agreement between researchers about the most important components which form the construct of mentalization. However, when strong associations cannot be explained by similar content, this may suggest that different measures are indeed capturing a latent construct of mentalization with separable components that are strongly related. First, the strong associations between the RFQ, MMQ, and MZQ may be explained by overlapping item content related to recognition and regulation of one's emotions. For instance, two of the three measures include items which capture awareness of one's emotions, particularly when affective experiences are intense (e.g., "Frequently it's

difficult for me to perceive my feelings at their full intensity" on the MZQ and "I always know what I feel" on the RFQ) and all three measures assess one's capacity for controlling their emotional states (e.g., "I sometimes feel like I am losing control of my emotions" on the MMQ, "Often I can't control my feelings" on the MZQ, and "When I get angry I say things that I later regret" on the RFQ). The very strong correlation between the two subscales with similar content, the Affect Regulation subscale on the MZQ and the Emotional Dyscontrol subscale on the MMQ (bivariate correlation is r = -.73) further supports this notion. Additionally, the moderate associations between the MentS with both the RFQ and MZQ, may be explained by similar shared content (e.g., "When I get upset, I am not sure whether I am sad, afraid, or angry" and "I can easily describe what I feel" on the MentS). Moreover, given that the items on the MentS_S in particular are indicative of identifying and regulating one's own emotions, this shared item content may be driving the associations between the MentS at the sum level with the RFQ, MZQ. Interestingly, the MentS_O also contains items about identifying others' emotions (e.g., "I can recognize other people's feelings") yet this subscale was weakly related with the RFQ and MZQ which both include items of a similar theme but reference one's *own* emotion identification. These findings may suggest that recognizing one's own versus other's emotions may be separable capacities that are not always related. Moreover, whether these findings signify that these measures are simply capturing emotion regulation skills or that mentalization is indeed a superordinate capacity with emotion regulation as an underlying component is unknown.

In addition to the strong association between the MMQ and the MZQ at the sum level, a strong link emerged between the Relational Discomfort subscale of the MMQ and the Refusing Self-Reflection subscale of the MZQ (r = -.60). This association may be explained by overlapping content in one item pertaining to discomfort in disclosing one's internal experiences

to others (e.g., “I am afraid to open up with other people” on the MMQ and “Most of the time I don’t feel like talking about my thoughts and feelings with others” on the MZQ). However, aside from this small similarity, the subscales appear to be capturing different content, specifically, avoiding thinking about affective experiences in the Refusing Self-Reflection subscale and feeling misunderstood and hurt by others in the Relational Discomfort subscale. As such, it is possible that these subscales were related because they are capturing different subcomponents that are both encapsulated within the broader construct of mentalizing. Another possibility for why these scales are strongly linked is that they both reflect components or consequences of avoidant attachment, which is characterized by avoidance of close, intimate relationships, discomfort with emotional vulnerability and refusing to rely on others (Mikulincer et al., 2003). Relatedly, low insight about one’s emotions may lead to difficulties empathizing with others’ emotional experiences and result in strained interpersonal relationships.

Aside from the overlapping content on the MentS and MZQ related to emotion identification and control, the MentS_S is also strongly related to the Refusing Self Reflection and the Psychic Equivalence subscales on the MZQ. These links may also be partially explained by similar content which specifies discomfort in disclosing one’s internal experiences to others (“While people talk about their feelings and needs my thoughts often drift away” on the MentS_S and “Most of the time I don’t feel like talking about my thoughts and feelings with others” on the MZQ). However, the former item on the MentS_S may be more related to disinterest in others’ internal states rather than discomfort expressing one’s own thoughts and feelings. There is also no clear overlap in item content across the MentS_S and the Psychic Equivalence subscale, perhaps suggesting that these subscales are indeed tapping different subcomponents of the latent construct of mentalization.

The weak associations between the IMQ with the MZQ at both the sum and subscale level may signify that the research teams who developed these measures are not in agreement about which psychological capacities are involved in mentalizing and relatedly, these tools may be measuring different, yet somewhat related latent constructs. Supporting this explanation, the IMQ and MZQ do not appear to share any item content and the purpose of developing each of these measures was markedly different. Hausberg et al. (2012) developed the MZQ for use with clinical populations in order to track their improvements in mentalization throughout treatment and thus, aimed to capture more severe deficits in mentalization. However, Wu et al. (2022) sought to develop a measure for non-clinical populations that would capture mentalization-based processes involved in social interactions, and specifically changes in these processes in different social circumstances. Even after considering the differences in the scope of each of the measures, a slightly stronger and moderate association would be expected between measures given that they are purportedly assessing the same latent construct.

Given that the IMQ and RFQ are related to a moderate degree and share some common content, this may signify some agreement between author teams in defining mentalization. This overlapping content is associated with insight into one's own actions (e.g., "I don't always know why I do what I do" on the RFQ and "I have accurate insight into why I act the way I do" on the IMQ) and others' thoughts (e.g., "People's thoughts are a mystery to me" on the RFQ and "I believe that I am good at telling what another person is thinking" on the IMQ). Interestingly, both the MMQ and the MentS appear to capture content related to one's understanding of their behaviors and the MMQ has items that are related to inferring others' thoughts. Unfortunately, the associations between the MMQ, MentS and IMQ could not be calculated in this study.

In sum, researchers' conceptualizations of what constitutes mentalization seem to converge for some subcomponents, specifically, emotion recognition and regulation, reflecting on and processing one's motivations for their behaviors and making accurate inferences about others' thoughts. Moreover, one indicator of the lower end of mentalizing abilities may be discomfort with disclosing one's internal states to others. Last, the IMQ and MZQ are suited for distinct populations and should not be used interchangeably.

Latent Factors Underlying Self-Report Measures

To gain further clarity on the boundaries that constrain this construct, the authors sought to extract the most fitting factor structure of the self-report measures. Based on the RFQ and MZQ, a 1-factor structure emerged however, most of the associations between the other measures and the extracted factor were low. The exceptions to these results were the MentS_S, Ego Strength, Relational Discomfort, and Emotional Dyscontrol subscales as well as the sum score on the MMQ, which had factor loadings of a similar strength to the RFQ and MZQ. These trends (e.g., higher associations with the MentS_S and some MMQ subscales and lower relations with the IMQ), align well with the results from the bivariate correlations. Although these results may suggest that there is one global all-encompassing capacity for mentalizing, this conclusion is in contrast to various theoretical assertions that mentalization is multidimensional and requires subordinate skills (e.g., emotion identification and control). The one factor structure also appears to be inconsistent with the broader conclusions suggested by the correlational work in the current study (e.g., that there are multiple subcomponents within the broader construct of mentalizing). However, it is possible that both the RFQ and MZQ are capturing only one of these subcomponents of mentalizing, specifically a subfactor related to identifying one's own emotions and affect regulation (given the overlapping content in this domain across those measures). Thus,

this particular finding of the one factor structure should be interpreted with caution and recognition of the limitations of this work discussed in detail below, namely that item level factor analyses were not possible with the current sample.

Convergence between Self-Report and Task-Based Measures of Mentalization

Moving beyond the associations between self-report measures, another aim of this study was to investigate the relations between self-report and task-based assessments. According to an overview by Luyten et al. (2012), the MASC represents a promising experimental/observational task to assess mentalization and prior research has supported construct validity of this measure in relation to mentalization theory. For instance, Fossati et al. (2018) found that 'no mentalizing' errors on the MASC (referring to situation-based inferences to explain individuals' behaviors) were positively related to various measures of BPD severity. This finding is consistent with the theoretical proposition that deficits in mentalization are a vulnerability factor in the development of BPD (Bateman & Fonagy, 2004b). Similarly, Sharp et al. (2011) found that hypermentalizing errors on the MASC were associated with BPD traits. However, only a few studies (Duval et al., 2018; Schwarzer et al., 2021a; Sharp et al., 2021; Rothschild-Yakar et al., 2019) have examined the convergent validity of the MASC with other measures of mentalizing and the results of these studies are mixed.

In contrast to predictions, relations between all self report measures and the MASC were very weak. The current study found a weaker correlation between the MZQ and MASC than Schwarzer et al. (2021a) and thus, did not replicate their findings. The extremely weak associations between the self-report measures and the MASC may be explained by advantages of behavioral measures and limitation of questionnaires. Reporting on one's own mentalization capacities likely requires moderate to high mentalizing skills and may introduce biases into the

data. For instance, an individual who often engages in hypermentalizing may think they are competent at inferring others' thoughts and feelings and their reports on a questionnaire will be in accordance with this belief. However, a task-based measure may reveal an individual's tendency towards over-interpreting mental states. Task-based and self-report measures may also have differences in the scope of mentalization they are capturing. For instance, only inferences about others' thoughts, feelings, and motivations are assessed in the MASC whereas questionnaires provide a more comprehensive assessment of one's mentalizing capacities related to both self and others which may not always be related. Some questionnaires (e.g., RFQ and MZQ) are predominantly focused on self-related mentalizing, however even the MentS which captures self- and other-related mentalizing was unrelated to MASC performance. Another potential explanation for the lack of convergence between the self report measures and the MASC is that the MASC is conducted in an emotionally neutral context where participants have no connection to the characters in which they are answering questions about. However, responding to questions on a self-report measure requires participants to reflect on their own experiences with mentalizing their own as well as others' internal states and as such, this task may elicit more prominent emotional reactions. A related key idea espoused by Mikolajczak and colleagues (2008) is that task-based measures may be indicative of one's capabilities within a certain domain whereas self-report measures demonstrate how that capability is translated and manifests within one's daily life. One implication of this notion is that self-report and task-based measures may be weakly associated.

In the regression analyses, none of the sum or subscale level variables of the self-report measures predicted performance on the MASC, with a few exceptions. At both the sum level and subscale level, the MentS acted as a suppressor variable. Specifically, when the MentS was

added to the model, the predictive power of the RFQ and MZQ increased and reversed the direction of the relationship between the RFQ and the MASC. The MentS also had a negative direct effect on MASC performance. Given that higher scores on each of these measures were indicative of adaptive/balanced mentalizing skills, it is unclear why the addition of the MentS brought about a negative association between the RFQ and performance on the MASC. Given that this is the first study to examine these measures within one statistical model, these novel and unusual findings must be replicated before conceptual explanations are put forward.

Revisiting the RFQ

The expected associations between the RFQ and MASC were not supported. The RFQ_U was not related to hypomentalizing errors and the RFQ_C was not associated with errors related to hypermentalizing. These null results aligned with Sharp et al.'s (2021) work but not Rothschild-Yakar et al.'s (2019) and Duval et al.'s (2018) findings. Given that these measures share a strikingly similar conceptual framework, the lack of convergence between them suggests that at least one of these tools lacks validity. While the MASC is a popular measure of mentalization that is still under investigation, there are many prominent psychometric issues related to the RFQ_U and RFQ_C as discussed by Müller et al. (2021) in detail. The analyses in the current study supported the notion that these subscales are capturing redundant information and do not represent distinct mentalization deficits. As such, future researchers who would like to use this measure should strictly employ the sum score as a measure of overall mentalizing. It is possible that the longer versions of the RFQ (46 and 54 item version) may adequately capture these distinct impairments. Unfortunately, most research has used the 8-item version (Müller et al., 2021).

Limitations and Future Directions

This research project represents the first endeavour to examine the convergence of multiple self-report and task-based measures of mentalizing and to investigate the best fitting factor structure. Despite the novelty of this research, there are noteworthy limitations. First, given that participants only completed one of the MMQ, MentS, or IMQ, the sample size for analyses including these measures were notably smaller than those involving the RFQ and MZQ which all participants completed. This practical constraint also prevented the researchers from calculating correlations between the MMQ, MentS, and IMQ and from using all five self-report measures to extract the factor structure. Additionally, the overall low sample size (N = 247) precluded the authors from conducting EFA at the item level however, this is an aim of future research after further data collection. In the testlet level EFA, items that had a similar theme, but were in different measures, could not aggregate together thereby limiting the authors' conclusions about the underlying factor structure of mentalization in the current study. It is anticipated that item-level analyses will more precisely elucidate the separable components of this broad construct. Another important limitation of this study was that the sample included non-clinical university students however, some measures were specifically designed for use in client samples with psychopathology (e.g., MZQ). Future work could aim to elucidate whether there are differences in the factor structure for clinical and non-clinical populations. Additionally, given that the construct of mentalization has started to gain recognition and is popular with English-speaking researchers and that some tools were validated in other languages (e.g., MZQ, MMQ, and MentS), another promising future direction is further validating the English versions of these tools. Moreover, the factor structure of all the self report measures used in this study remains uncertain for one of two reasons, 1) there have been inconsistent findings

regarding the dimensional structure (e.g., RFQ and MZQ) or 2) no additional studies besides the seminal research have investigated the structure (e.g., IMQ and MMQ). Furthermore, given the critiques of the RFQ by Müller et al. (2021), further work should aim to replicate their factor analytic investigation of two types of scoring methods (e.g., based on the original RFQ_C and RFQ_U scoring methods and scoring protocols to generate a sum score). Last, an important limitation of this study is that the interview based Reflective Functioning Scale (RFS) which represents the gold standard measure of mentalization was not included. Practical constraints such as the cost and the time-intensive nature of learning the coding system, hiring research assistants to conduct interviews, transcribing, and coding, prevented the researchers from employing this tool in this study.

This work represents a significant step towards exploring convergent validity of mentalization measures and establishing the conceptual boundaries in which mentalization is situated. As future work moves forward with further specifying this construct, it is vital for researchers to be remain open-minded about this process and make active efforts to avoid "generative entrenchment" (Wimsatt, 1986). This phenomenon occurs when concepts become well-established within the literature to the point where other theories and constructs depend on them and it becomes increasingly difficult to challenge the underlying concepts, even if they are dubious. Indeed, "conceptual clarification and construct validation should be seen as important and valuable parts of research, and validation should be taken to be an iterative and ongoing process instead of just a hurdle that needs be crossed" (Eronen & Bringmann, 2021, p. 785).

Tables

Table 1

Demographic Information

Characteristic	Percentage (%)
Gender	
Man/Transman	22.36
Woman/Transwoman	77.02
Gender queer/non-conforming /non-binary	.62
Sex	
Male	22.36
Female	77.64
Intersex	0
Primary Racial Group	
South Asian/Southeast Asian/East Asian/ Middle Eastern	45.96
Black/African	3.73
Hispanic	1.86
Indigenous (First Nations, Métis, or Inuit)	.62
White	4.37
Mixed	3.11
West Indian	1.24
Other	.62
Prefer not to answer	2.49
Birth Year	
1979-1999	13.67
2000-2002	57.76
2003 or later	27.95
Prefer not to answer	.62

Note. N = 161

Table 2

Bivariate Correlations between MASC Performance and Sum Level Scores on Self-Report Measures

Measure	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.
1. MASC Correct	-										
2. MASC Hyper	-.303**	-									
3. MASC Hypo	-.357**	.179**	-								
4. MASC No Ment	-.293**	.073	.363**	-							
5. RFQ Sum	-.012	-.069	-.065	-.13*	**.80**						
6. RFQ_C	-.011	-.051	-.079	-.156*	.948**	**.73**					
7. RFQ_U	.023	.090	.065	.095	-.947**	-.861**	**.75**				
8. MZQ Sum	.111	-.109	-.163*	-.078	.622**	.541**	-.596**	**.79**			
9. IMQ Sum	.194	-.211	-.097	-.048	.466**	.415**	-.434**	.33**	**.71**		
10. MMQ Sum	.127	-.003	-.128	-.141	.649**	.588**	-.598**	.663**	X	**.85**	

11. MentS Sum	-.047	.066	.066	-.103	.458**	.391**	-.421**	.51**	X	X	**.85**

Note. X signifies that correlations could not be calculated because participants completed either the IMQ, MentS, or MMQ. MASC Hyper refers to hypermentalizing errors on the MASC. MASC Hypo refers to hypomentalizing errors on the MASC. MASC No Ment refers to errors on the MASC which reflect fact or situation-based inferences to explain characters' behaviors and makes no reference to their mental states. Cronbach's alpha coefficient is in bold along the diagonal.

* $p<.05$
** $p<.01$

Table 3

Bivariate Correlations between the RFQ and Subscale Level Scores on the MMQ and MZQ

Measure	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.
1. RFQ Sum	**.80**										
2. MMQ Reflexivity	.207	**.78**									
3. MMQ Ego Strength	.583**	.377**	**.82**								
4. MMQ Relational Attunement	.420**	.436**	.449**	**.70**							
5. MMQ Relational Discomfort	-.455*	-.104	-.392**	-.222*	**.70**						
6. MMQ Distrust	-.178	.039	-.063	-.070	.411**	**.64**					
7. MMQ Emotion Dyscontrol	-.587**	-.057	-.438**	-.006	.424**	.213	**.66**				
8. MZQ Refusing SR	.311**	.253*	.245*	.168	-.599**	-.338**	-.310**	**.53**			

9. MZQ Emotion Awareness	.559**	.177	.408**	.187	-.317**	-.125	-.457**	.462**	**.64**		
10. MZQ Psychic Equivalence	.481**	-.049	.364**	.009	-.464**	-.237*	-.396**	.343**	.393**	**.61**	
11. MZQ Affect Regulation	.519**	.073	.442**	.158	-.478**	-.216	-.729**	.353**	.423**	.449**	**.47**

Table 4

Bivariate Correlations between MASC Performance and Subscale Level Scores on the RFQ, IMQ, and MentS

Measure	1.	2.	3.	4.	5.	7.	8.	9.	10.	11.	12.
1. MASC Correct	-										
2. MASC Hyper	-.303**	-									
3. MASC Hypo	-.357**	.179**	-								
4. MASC No Ment	-.293**	.073	.363**	-							
5. RFQ Sum	-.012	-.069	-.036	-.130*	**.73**						
7. IMQ Perspective	.11	-.038	-.100	-.081	.290*	**.71**					
8. IMQ Metacog	.03	-.138	.082	.018	.321*	.426**	**.63**				
9. IMQ Meta-ment	.272*	-.254*	-.202	-.048	.338**	.106	.018	**.58**			
10. MentS_S	.121	-.029	-.071	-.266*	.544**	X	X	X	**.80**		

11. MentS_O	-.160	.147	.142	.031	.219	X	X	X	.309**	**.81**	
12. MentS_M	-.113	.060	.105	.059	.225*	X	X	X	.320**	.553**	**.62**

Note. X signifies that correlations could not be calculated because participants completed either the IMQ, MentS, or MMQ. MASC Hyper refers to hypermentalizing errors on the MASC. MASC Hypo refers to hypomentalizing errors on the MASC. MASC No Ment refers to errors on the MASC which reflect fact or situation-based inferences to explain characters' behaviors and makes no reference to their mental states. IMQ Perspective refers to the perspective-taking subscale (e.g., mentalizing about others). IMQ Metacog refers to the metacognition subscale (e.g., mentalizing about one's self) and IMQ Meta-ment refers to the meta-mentalization subscale. MentS_S refers to self-related mentalization subscale, MentS_O refers to other-related mentalization subscale, and MentS_M refers to the motivation to mentalize subscale. Cronbach's alpha coefficient is in bold along the diagonal.

* $p<.05$
** $p<.01$

Table 5

Bivariate Correlations between Subscale Level Scores on the MZQ and IMQ

Measure	1	2.	3.	4.	5.	6.	7.
1. MZQ Refusing SR	**.53**						
2. MZQ Emotion Awareness	.462**	**.64**					
3. MZQ Psychic Equivalence	.343**	.393**	**.61**				
4. MZQ Affect Regulation	.353**	.423**	.449**	**.47**			
5. IMQ Perspective	-.033	.104	.049	.220*	**.71**		
6. IMQ Metacog	.157	.263*	.269*	.306**	.426**	**.63**	
7. IMQ Meta-ment	.117	.105	.265*	.226*	.106	.018	**.58**

Note. MZQ Refusing SR refers to the Refusing Self-Reflection subscale. MentS_S refers to self-related mentalization subscale, MentS_O refers to other-related mentalization subscale, and MentS_M refers to the motivation to mentalize subscale. IMQ Perspective refers to the perspective-taking subscale (e.g., mentalizing about others). IMQ Metacog refers to the metacognition subscale (e.g., mentalizing about one's self) and IMQ Meta-ment refers to the meta-mentalization subscale. Cronbach's alpha coefficient is in bold along the diagonal.

* $p<.05$
** $p<.01$

Table 6

Bivariate Correlations between MASC Performance and Subscale Level Scores on the MentS and MZQ

Measure	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.
1. MASC Correct	-										
2. MASC Hyper	-.303**	-									
3. MASC Hypo	-.357**	.179**	-								
4. MASC No Ment	-.293**	.073	.363**	-							
5. MentS_S	.121	-.029	-.071	-.266*	**.80**						
6. MentS_O	-.16	.147	.142	.031	.309**	**.81**					
7.MentS_M	-.113	.060	.105	.059	.320**	.553**	**.62**				
8. MZQ Refusing SR	.165*	-.127	-.232**	-.121	.558**	.032	.173	**.53**			
9. MZQ Emotional Awareness	.032	-.042	-.036	-.034	.609**	.212	.075	.462**	**.64**		

10. MZQ Psychic Equivalence	.053	-.074	-.058	-.032	.645**	.228	.247*	.343**	.393**	**.61**	
11. MZQ Affect Regulation	.079	-.080	-.171**	-.043	.450**	-.052	-.034	.353**	.423**	.449**	**.47**

Note. MASC Hyper refers to hypermentalizing errors on the MASC. MASC Hypo refers to hypomentalizing errors on the MASC. MASC No Ment refers to errors on the MASC which reflect fact or situation based inferences to explain characters' behaviors and makes no reference to their mental states. MZQ Refusing SR refers to the Refusing Self-Reflection subscale. MentS_S refers to self-related mentalization subscale, MentS_O refers to other-related mentalization subscale, and MentS_M refers to the motivation to mentalize subscale. Cronbach's alpha coefficient is in bold along the diagonal.

* $p<.05$
** $p<.01$

Table 7

Bivariate Correlations between MASC Performance and Subscale Level Scores on the MMQ

Measure	1	2.	3.	4.	5.	6.	7.	8.	9.	10.
1. MASC Correct	-									
2. MASC Hyper	-.303**	-								
3. MASC Hypo	- .357**	.179**	-							
4. MASC No Ment	-.293**	.073	.363**	-						
5. MMQ Reflexivity	.068	-.020	.036	-.168	**.78**					
6. MMQ Ego Strength	-.025	.145	-.055	-.077	.377**	**.82**				
7. MMQ Relational Attunement	.055	.031	-.083	-.079	.436**	.449**	**.70**			
8. MMQ Relational Discomfort	-.189	.073	.244*	.052	-.104	-.392**	-.222*	**.70**		
9. MMQ Distrust	-.187	.117	.102	.161	.039	-.063	-.070	.411**	**.64**	

10. MMQ Emotion Dyscontrol	.001	-.007	.030	-.032	-.057	-.438**	-.006	.424**	.213	**.66**

Note. MASC Hyper refers to hypermentalizing errors on the MASC. MASC Hypo refers to hypomentalizing errors on the MASC. MASC No Ment refers to errors on the MASC which reflect fact or situation based inferences to explain characters' behaviors and makes no reference to their mental states. Cronbach's alpha coefficient is in bold along the diagonal.

* $p<.05$
** $p<.01$

Table 8

Self-Report Measures at the Sum Level as Predictors of Performance on the MASC

Predictor	β	SE	R^2	F change (df)	*p*
Dependent variable: MASC Correct					
Model 1: RFQ, MZQ		4.87	.014	1.66 (2,236)	.193
RFQ	.001	.067			.994
MZQ	.117	.044			.156
Model 2: RFQ, MZQ, and MMQ		4.33	.021	.549(3,76)	.65
Model 3: RFQ, MZQ, and IMQ		3.91	.096	2.75 (3,78)	.048
RFQ	.121	.101			.424
MZQ	.170	.064			.232
IMQ	.081	.064			.506
Model 4: RFQ, MZQ, and MentS		4.57	.128	3.43(3,70)	.021
RFQ	-.344	-.344			.016
MZQ	.422	.422			.005
MentS	-.105	-.105			.434

Note. All variables are the sum level scores of the self-report measures.

Table 9

Bivariate and Partial Correlations between RFQ, MZQ, and MASC Performance

	MASC Correct (partial correlation)	MASC Correct (zero-order correlation)
RFQ Sum	-.130	-.012
MZQ Sum	.227	.111

Note. The second column represents the partial correlations between the RFQ and MASC performance and between the MZQ and MASC performance, when controlling for the sum score on the MentS. The third column represents the bivariate correlations between the RFQ, MZQ, and MASC performance.

Table 10

Self-Report Measures at the Subscale Level as Predictors of Performance on the MASC

Predictor	β	SE	R^2	F change (df)	*p*
Dependent variable: MASC Correct					
Model 1: RFQ Sum four MZQ subscales		4.85	.034	1.63(5,233)	.152
RFQ Sum	.046	.070			.593
MZQ Refusing SR	.169	.114			.025
MZQ Emotion Awareness	-.100	.127			.238
MZQ Psychic Equivalence	-.010	.116			.901
MZQ Affect Regulation	.068	.162			.397
Model 2: Model 1 and all three IMQ subscales		3.84	.184	2.05 (8,73)	.052
Model 3: Model 1 and all six MMQ subscales		4.36	.112	.78(11,68)	.656
Model 4: Model 1 and all three MentS subscales		4.55	.203	2.04(8,64)	.055
Model 5: Model 1 and IMQ Perspective Taking		3.87	.150	2.11 (6,75)	.051
Model 6: Model 1 and IMQ Metacognition		3.85	.157	2.33 (6,75)	.040
RFQ Sum	.186	.102			.227
MZQ Refusing SR	.175	.149			.170

MZQ Emotion Awareness	-.209	.179			.155
MZQ Psychic Equivalence	.105	.159			.406
MZQ Affect Regulation	.230	.247			.089
IMQ Metacognition	-.101	.113			.381
Model 7: Model 1 and IMQ Meta-mentalization		3.82	.172	2.61 (6,75)	.024
RFQ Sum	.117	.104			.456
MZQ Refusing SR	.166	.148			.187
MZQ Emotion Awareness	-.186	.179			.204
MZQ Psychic Equivalence	.070	.158			.574
MZQ Affect Regulation	.207	.243			.120
IMQ Meta-mentalization	.167	.130			.146
Model 8: Model 1 and MMQ Reflexivity		4.34	.052	.67(6,73)	.671
Model 9: Model 1 and MMQ Ego Strength		4.32	.062	.80 (6,73)	.572
Model 10: Model 1 and MMQ Relational Attunement		4.34	.055	.71 (6,73)	.641
Model 11: Model 1 and MMQ Relational Discomfort		4.30	.073	.96 (6,73)	.457
Model 12: Model 1 and MMQ Distrust		4.28	.079	1.04 (6,73)	.407
Model 13: Model 1 and MMQ Emotional Dyscontrol		4.34	.053	.68(6,73)	.666

Model 14: Model 1 and MentS_S		4.50	.191	2.64 (6,67)	.023
RFQ Sum	-.361	.125			.016
MZQ Refusing SR	.208	.222			.153
MZQ Emotion Awareness	-.132	.226			.380
MZQ Psychic Equivalence	-.005	.253			.977
MZQ Affect Regulation	.300	.288			.040
MentS_S	.148	.139			.400
Model 15: Model 1 and MentS_O		4.53	.187	2.53(6,67)	.029
RFQ Sum	-.321	.124			.031
MZQ Refusing SR	.258	.207			.058
MZQ Emotion Awareness	-.083	.223			.576
MZQ Psychic Equivalence	.055	.242			.728
MZQ Affect Regulation	.265	.301			.081
MentS_O	-.066	.117			.584
Model 16: Model 1 and MentS_M		4.56	.192	2.66 (6,67)	.022
RFQ Sum	-.308	.124			.038
MZQ Refusing SR	.281	.210			.042
MZQ Emotion Awareness	-.120	.221			.414
MZQ Psychic Equivalence	.082	.244			.619
MZQ Affect Regulation	.254	.301			.094
MentS_M	-.108	.139			.372

Table 11

Bivariate and Partial Correlations between RFQ, MZQ, Subscales and MASC Performance

	MASC Correct (partial correlation)	MASC Correct (zero-order correlation)
RFQ Sum		-.012
Controlling for MentS_S	-.250	
Controlling for MentS_O	-.104	
Controlling for MentS_M	-.109	
MZQ Affect Regulation	.188	.079
Controlling for MentS_S		
MZQ Refusing Self Reflection	.292	.165
Controlling for MentS_M		

Note. The second column represents the partial correlations between the RFQ and MASC performance and between the MZQ and MASC performance, when controlling for the MentS. The third column represents the bivariate correlations between the RFQ, MZQ, and MASC performance

Table 12

Rotated Factor Matrix from EFA with RFQ and MZQ Testlets

Testlet	Loadings onto Factor 1	Loadings onto Factor 2
RFQ Certainty 1	.45	.40
RFQ Certainty 2	.93	.28
RFQ Uncertainty 1	-.39	-.57
RFQ Uncertainty 2	-.92	-.29
MZQ Refusing Self Reflection	.12	.52
MZQ Emotional Awareness	.26	.68
MZQ Psychic Equivalence	.24	.58
MZQ Affect Regulation	.24	.64

Note. Extraction method: principal axis factoring analysis; Rotation method: Varimax with Kaiser normalization.

Table 13

Factor Matrix from EFA after removing RFQ Testlet

Testlet	Loadings onto Factor 1
RFQ Certainty 1	.60
RFQ Certainty 2	.68
RFQ Uncertainty 1	-.71
MZQ Refusing Self Reflection	.48
MZQ Emotional Awareness	.70
MZQ Psychic Equivalence	.61
MZQ Affect Regulation	.67

Subscale/Sum	Bivariate Correlations
IMQ Sum	.447**
IMQ Others	.215
IMQ Self	.371**
IMQ Meta-Ment	.316**
MMQ Sum	.718**
MMQ Reflexivity	.188
MMQ Ego Strength	.625**
MMQ Relational Attunement	.358**
MMQ Relational Discomfort	-.573**
MMQ Distrust	-.24*
MMQ Emotional Dyscontrol	-.70**
MentS Sum	.52**
MentS Self	.711**
MentS Others	.203
MentS Motivation	.179

* $p<.05$
** $p<.01$

References

Abate, A., Marshall, K., Sharp, C., & Venta, A. (2017). Trauma and aggression: Investigating the mediating role of mentalizing in female and male inpatient adolescents. *Child Psychiatry and Human Development, 48*(6), 881–89. https://doi.org/1.1007/s10578-017-0711-6

Ahmadian, Z., & Ghamarani, A. (2021). Reliability and validity of Persian version of the Mentalization Scale in university students. *Journal of Fundamentals of Mental Health, 23*(4): 233-24.

Ainsworth, M. D. S., Blehar, M. C., Waters, E., & Wall, S. (1978). *Patterns of attachment: A psychological study of the strange situation*. Taylor & Francis Group. https://doi.org/1.4324/9780203758045

Allen, J. G. (2006). Mentalizing in practice. In J. G. Allen & P. Fonagy (Eds.), *Handbook of mentalization-based treatment* (pp. 3–30). John Wiley & Sons Ltd.

Anis, L., Perez, G., Benzies, K. M., Ewashen, C., Hart, M., & Letourneau, N. (2020). Convergent validity of three measures of reflective function: Parent Development Interview, Parental Reflective Function Questionnaire, and Reflective Function Questionnaire. *Frontiers in Psychology, 11*, 574719. https://doi.org/1.3389/fpsyg.202.574719

Antonsen, B. T., Johansen, M. S., Rø, F. G., Kvarstein, E. H., & Wilberg, T. (2016). Is reflective functioning associated with clinical symptoms and long-term course in patients with personality disorders? *Comprehensive Psychiatry, 64*, 46-58. https://doi.org/1.1016/j.comppsych.2015.05.016

Apperly, I. A. (2012). What is “theory of mind”? Concepts, cognitive processes and individual differences. *Quarterly Journal of Experimental Psychology*, *65*(5), 825–839. https://doi.org/1.1080/17470218.2012.676055

Arnott, B., & Meins, E. (2007). Links among antenatal attachment representations, postnatal mind-mindedness, and infant attachment security: A preliminary study of mothers and fathers. *Bulletin of the Menninger Clinic, 71*(2), 132–149. https://doi.org/1.1521/bumc.2007.71.2.132

Badoud, D., Luyten, P., Fonseca-Pedrero, E., Eliez, S., Fonagy, P., & Debbané, M. (2015) The French Version of the Reflective Functioning Questionnaire: Validity Data for Adolescents and Adults and Its Association with Non-Suicidal Self-Injury. *PLoS ONE, 10*(12): e0145892. https://doi.org/1.1371/journal.pone.0145892

Ballespí, S., Vives, J., Sharp, C., Tobar, A., & Barrantes-Vidal, N. (2019). Hypermentalizing in social anxiety: Evidence for a context-dependent relationship. *Frontiers in Psychology, 10*, 1501. https://doi.org/1.3389/fpsyg.2019.01501

Barberis, N., Cannavò, M., Calaresi, D., & Verrastro, V. (2022). Reflective functioning and alexithymia as mediators between attachment and psychopathology symptoms: Cross-sectional evidence in a community sample. *Psychology, Health & Medicine*, 1–12. https://doi.org/1.1080/13548506.2022.2045331

Baron-Cohen, S., & Wheelwright, S. (2004). The Empathy Quotient: An investigation of adults with Asperger syndrome or high functioning autism and normal sex differences. *Journal of Autism and Developmental Disorders, 34*, 163–175. https://doi.org/1.1023/B:JADD.0000022607.19833.00

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001a). The "Reading the mind in the eyes" Test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry, 42*(2), 241–251. https://doi.org/1.1111/1469-761.00715

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001b). The autism spectrum quotient (AQ): evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism Developmental Disorders, 31*(1), 5-17. https://doi.org/1.1023/a:1005653411471

Bartholomew, K., & Horowitz, L. M. (1991). Attachment styles among young adults: A test of a four-category model. *Journal of Personality and Social Psychology, 61*(2), 226–244. https://doi.org/1.1037/0022-3514.61.2.226

Bartlett, M. S. (1951). The effect of standardization on a Chi-square approximation in factor analysis. *Biometrika, 38*(3/4), 337-344. https://doi.org/1.1093/biomet/38.3-4.337

Bateman, A., Campbell, C., Luyten, P., & Fonagy, P. (2018). A mentalization-based approach to common factors in the treatment of borderline personality disorder. *Current Opinion in Psychology*, *21*, 44-49. https://doi.org/1.1016/j.copsyc.2017.09.005

Bateman, A., & Fonagy, P. (2004a). *Psychotherapy for Borderline Personality Disorder: Mentalization Based Treatment.* Oxford: Oxford University Press.

Bateman, A., & Fonagy, P. (2004b). Mentalization-based treatment of BPD. *Journal of Personality Disorders*, *18*, 36–51.

Bateman, A., & Fonagy, P. (2012). *Handbook of mentalizing in mental health practice*. American Psychiatric Publishing, Inc.

Baer, R. A., Smith, G. T., Hopkins, J., Krietemeyer, J., & Toney, L. (2006). Using Self-Report Assessment Methods to Explore Facets of Mindfulness. *Assessment, 13*(1), 27–45. https://doi.org/1.1177/1073191105283504

Beaulieu-Pelletier, G., Bouchard, M.-A., & Philippe, F. L. (2013). Mental States Task (MST): Development, validation, and correlates of a self-report measure of mentalization. *Journal of Clinical Psychology, 69*(7), 671–695. https://doi.org/1.1002/jclp.21942

Belvederi Murri, M., Ferrigno, G., Penati, S., Muzio, C., Piccinini, G., Innamorati, M., Ricci, F., Pompili, M., & Amore, M. (2017). Mentalization and depressive symptoms in a clinical sample of adolescents and young adults. *Child and Adolescent Mental Health, 22*(2), 69–76. https://doi.org/1.1111/camh.12195

Benoit, A. M. (2020). *Examining relationships between early childcare teachers' adult attachment orientations and quality of interaction in the infant classroom* (5245). [Doctoral dissertation, Louisiana State University and Agricultural and Mechanical College]. LSU Doctoral Dissertations.

Berthelot, N., Ensink, K., Bernazzani, O., Normandin, L., Luyten, P., & Fonagy, P. (2015). Intergenerational transmission of attachment in abused and neglected mothers: The role of trauma-specific reflective functioning. *Infant Mental Health Journal, 36*(2), 200–212. https://doi.org/1.1002/imhj.21499

Berthelot, N., Lemieux, R., Garon-Bissonnette, J., Lacharité, C., & Muzik, M. (2019). The The protective role of mentalizing: Reflective functioning as a mediator between child maltreatment, psychopathology and parental attitude in expecting parents. *Child Abuse & Neglect, 95*, 104065. https://doi.org/1.1016/j.chiabu.2019.104065

Bhola, P., & Mehrotra, K. (2021). Associations between countertransference reactions towards patients with borderline personality disorder and therapist experience levels and mentalization ability. *Trends in Psychiatry and Psychotherapy,43*(2),116-125. http://dx.doi.org/1.47626/2237-6089-2020-0025

Bizzi, F., Riva, A., Borelli, J. L., Charpentier-Mora, S., Bomba, M., Cavanna, D., & Nacinovich, R. (2022). The Italian version of the Reflective Functioning Questionnaire: Validity within a sample of adolescents and associations with psychological problems and alexithymia. *Journal of Clinical Psychology, 78*(4), 503–516. https://doi.org/1.1002/jclp.23218

Bouchard, M. A., Audet, C., Picard, C., Carrier, M., & Milcent, M. P. (2001). *The Mental States Rating System. Scoring manual.* Unpublished manuscript. Psychology Department, University of Montreal.

Bouchard, M. A., Target, M., Lecours, S., Fonagy, P., Tremblay, L.-M., Schachter, A., & Stein, H. (2008). Mentalization in adult attachment narratives: Reflective functioning, mental states, and affect elaboration compared. *Psychoanalytic Psychology, 25*(1), 47–66. https://doi.org/1.1037/0736-9735.25.1.47

Bressi, C., Fronza, S., Minacapelli, E., Nocito, E. P., Dipasquale, E., Magri, L., Lionetti, F., & Barone, L. (2016). Short-term psychodynamic psychotherapy with mentalization-based techniques in major depressive disorder patients: Relationship among alexithymia, reflective functioning, and outcome variables-A pilot study. *Psychology and Psychotherapy: Theory, Research and Practice, 90*(3), 299–313. https://doi.org/1.1111/papt.12110

Brugnera, A., Zarbo, C., Compare, A., Talia, A., Tasca, G. A., de Jong, K., Greco, A., Greco, F.,

Pievani, L., Auteri, A., & Lo Coco, G. (2020). Self- reported reflective functioning mediates the association between attachment insecurity and well-being among psychotherapists. *Psychotherapy Research, 31*(2), 247-257. https://doi.org/1.1080/10503307.202.1762946

Calaresi, D., & Barberis, N. (2019). The relationship between reflective functioning and alexithymia. *Journal of Clinical and Developmental Psychology, 1*(2), 12-23.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*(2), 245–276. https://doi.org/1.1207/s15327906mbr0102_10

Chiesa, M., Luyten, P., & Fonagy, P. (2021). Two-year follow-up and changes in reflective functioning in specialist and nonspecialist treatment models for personality disorder. *Personality Disorders: Theory, Research, and Treatment, 12*(3), 249–26. https://doi.org/1.1037/per0000464

Choi-Kain, L. W., & Gunderson, J. G. (2008). Mentalization: Ontogeny, assessment, and application in the treatment of borderline personality disorder. *The American Journal of Psychiatry, 165*(9), 1127–1135. https://doi.org/1.1176/appi.ajp.2008.07081360

Chow, C.-C., Nolte, T., Cohen, D., Fearon, R. M. P., & Shmueli-Goetz, Y. (2017). Reflective functioning and adolescent psychological adaptation: The validity of the Reflective Functioning Scale–Adolescent Version. *Psychoanalytic Psychology, 34*(4), 404–413. https://doi.org/1.1037/pap0000148

Chrysikou, E. G., & Thompson, W. J. (2016). Assessing cognitive and affective empathy through the Interpersonal Reactivity Index: An argument against a two-factor model. *Assessment, 23*(6), 769–777. https://doi.org/1.1177/1073191115599055

Cortés-Garcia, L., McLaren, V., Vanwoerden, S., & Sharp, C. (2021). Attachment, mentalizing, and eating disorder symptoms in adolescent psychiatric inpatients and healthy controls: A test of a mediational model. *Eating and Weight Disorders, 26*(4), 1159–1168. https://doi.org/1.1007/s40519-020-01017-z

Cosenza, M., Ciccarelli, M., & Nigro, G. (2019). The steamy mirror of adolescent gamblers: Mentalization, impulsivity, and time horizon. *Addictive Behaviors, 89,* 156–162. https://doi.org/1.1016/j.addbeh.2018.1.002

Cucchi, A., Hampton, J. A., & Moulton-Perkins, A. (2018). Using the validated Reflective Functioning Questionnaire to investigate mentalizing in individuals presenting with eating disorders with and without self-harm. *PeerJ, 6,* e5756. https://doi.org/1.7717/peerj.5756

Daley, A. E. (2014). *Reflective functioning and differentiation-relatedness during pregnancy and infant attachment outcomes at one year*. [Doctoral dissertation, the City University of New York]. Cuny Academic Works.

Daudert, E. (2002). The Reflective Self-Functioning Scale. In B. Strauss, H. Kaechele, & A. Buchheim (Eds.), *Clinical attachment research: Theory, methods, results* (pp. 54 - 67). Stuttgart: Schattauer Publishers.

Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology, 10,* 85.

Davis, M. H. (1983). Measuring individual-differences in empathy - Evidence for a multidimensional approach. *Journal of Personality and Social Psychology, 44*(1), 113-126. https://doi.org/1.1037/0022-3514.44.1.113

Decety, J., & Jackson, P. L. (2004). The functional architecture of human empathy. *Behavioral and Cognitive Neuroscience Reviews, 3*(2), 71–10.

https://doi.org/1.1177/1534582304267187

Derogatis, L. R. (1993). BSI Brief Symptom Inventory: Administration, scoring, and procedure manual (4th Ed.). Minneapolis, MN: National Computer Systems.

De Meulemeester, C., Vansteelandt, K., Luyten, P., & Lowyck, B. (2018). Mentalizing as a mechanism of change in the treatment of patients with borderline personality disorder: A parallel process growth modeling approach. *Personality Disorders: Theory, Research, and Treatment, 9*(1), 22–29. https://doi.org/1.1037/per0000256

Diamond, D., Levy, K. N., Clarkin, J. F., Fischer-Kern, M., Cain, N. M., Doering, S., Hörz, S., & Buchheim, A. (2014). Attachment and mentalization in female patients with comorbid narcissistic and borderline personality disorder. *Personality Disorders: Theory, Research, and Treatment, 5*(4), 428–433. https://doi.org/1.1037/per0000065

Dimitrijević, A., Hanak, N., Altaras Dimitrijević, A., & Jolić Marjanović, Z. (2018). The Mentalization Scale (MentS): A self-report measure for the assessment of mentalizing capacity. *Journal of Personality Assessment, 100*(3), 268–28. https://doi.org/1.1080/00223891.2017.1310730

Dimitrijević, A., Ristein, L., Licata, M., Hamburger, A., Altaras Dimitrijević, A., & Jolić Marjanović, Z. (2017). *Validation of the German version of the Mentalization Scale (D-MentS).* Unpublished manuscript.

Dinsdale, N., & Crespi, B. J. (2013). The borderline empathy paradox: Evidence and conceptual models for empathic enhancements in borderline personality disorder. *Journal of Personality Disorders, 27*(2), 172–195. https://doi.org/1.1521/pedi_2012_26_071

Djordjevic, T., & Dordević, M. M. (2019). Recognizing emotions, attachment and mentalization capacity. *International Journal of Education and Psychology in the Community, 9* (1 & 2), 7–26.

Drogar, E., Fathi-Ashtiani, A., & Ashrafi, E. (2020). Validation and reliability of the Persian version of the Mentalization Questionnaire. *Journal of Clinical Psychology, 12* (45), 1-12. https://doi.org/1.22038/JFMH.2021.18969

Duval, J., Ensink, K., Normandin, L., Sharp, C., & Fonagy, P. (2018). Measuring reflective functioning in adolescents: Relations to personality disorders and psychological difficulties. *Adolescent Psychiatry, 8*(1), 5-2.

Dziobek, I., Fleck, S., Kalbe, E., Rogers, K., Hassenstab, J., Brand, M., Kessler, J., Woike, J. K., Wolf, O. T., & Convit, A. (2006). Introducing MASC: A Movie for the Assessment of Social Cognition. *Journal of Autism and Developmental Disorders, 36*(5), 623–636. https://doi.org/1.1007/s10803-006-0107-0

Ekeblad, A., Falkenström, F., & Holmqvist, R. (2016). Reflective functioning as predictor of working alliance and outcome in the treatment of depression. *Journal of Consulting and Clinical Psychology*, *84*(1), 67–78. https://doi-org.1.1037/ccp0000055

Eloranta, S. J., Kaltiala, R., Lindberg, N., Kaivosoja, M., & Peltonen, K. (2020). Validating measurement tools for mentalization, emotion regulation difficulties and identity diffusion among Finnish adolescents. *Nordic Psychology*. https://doi.org/1.1080/19012276.202.1863852

Ensink, K. (2004). *Assessing theory of mind, affective understanding and reflective functioning in primary school-aged children*. [Doctoral dissertation, University of London]. UCL Discovery.

Ensink, K., Bégin, M., Normandin, L., & Fonagy, P. (2016). Maternal and child reflective functioning in the context of child sexual abuse: Pathways to depression and externalising difficulties. *European Journal of Psychotraumatology, 7*(1), 30611. https://doi.org/1.3402/ejpt.v7.30611

Ensink, K., Begin, M., Normandin, L., Godbout, N., & Fonagy, P. (2017). Mentalization and dissociation in the context of trauma: Implications for child psychopathology. *Journal of Trauma & Dissociation, 18*(1), 11–3. https://doi.org/1.1080/15299732.2016.1172536.

Ensink, K., Berthelot, N., Bernazzani, O., Normandin, L., & Fonagy, P. (2014). Another step closer to measuring the ghosts in the nursery: Preliminary validation of the Trauma Reflective Functioning Scale. *Frontiers in Psychology, 5*, 1471. https://doi.org/1.3389/fpsyg.2014.01471

Ensink, K., Maheux, J., Normandin, L., Sabourin, S., Diguer, L., Berthelot, N., & Parent, K. (2013). The impact of mentalization training on the reflective function of novice therapists: A randomized controlled trial. *Psychotherapy Research, 23*(5), 526–538. https://doi.org/1.1080/10503307.2013.800950

Ensink, K., Normandin, L., Target, M., Fonagy, P., Sabourin, S., & Berthelot, N. (2015). Mentalization in children and mothers in the context of trauma: An initial study of the validity of the Child Reflective Functioning Scale. *British Journal of Developmental Psychology, 33*(2), 203–217. https://doi.org/1.1111/bjdp.12074

Epstein, S. (1983). *Mother-Father-Peer Scale*. Unpublished manuscript. University of Massachusetts, Amherst.

Eronen, M. I., & Bringmann, L. F. (2021). The theory crisis in psychology: How to move forward. *Perspectives on Psychological Science, 16*(4), 779–788. https://doi.org/1.1177/1745691620970586

Euler, S., Hüwe, L., Gablonski, T. C., Dehoust, M., Schulz, H., Brütt, A. L., & Andreas, S. (2022). Mentalizing Mediates the Association between Narcissism and Psychotherapeutic Treatment Outcome in a Mixed Clinical Sample. *Psychopathology*, 1-1. https://doi-org.1.1159/000524203

Eysenck, H., & Eysenck, F. (1975). Manual of the Eysenck Personality Questionnaire. London: Hodder and Stoughton.

Falkenström, F., Solbakken, O. A., Möller, C., Lech, B., Sandell, R., & Holmqvist, R. (2014). Reflective functioning, affect consciousness, and mindfulness: Are these different functions? *Psychoanalytic Psychology, 31*(1), 26–4. https://doi.org/1.1037/a0034049

Fekete, K., Török, E., Makkos, Z., Kelemen, O., Csigó, K., & Kéri, S. (2020). Mentalization across the psychosis spectrum. *Schizophrenia Research, 215*, 471–472. https://doi.org/1.1016/j.schres.2019.08.023

Fischer-Kern, M., Buchheim, A., Hörz, S., Schuster, P., Doering, S., Kapusta, N. D., Taubner, S., Tmej, A., Rentrop, M., Buchheim, P., & Fonagy, P. (2010). The relationship between personality organization, reflective functioning, and psychiatric classification in borderline personality disorder. *Psychoanalytic Psychology*, *27*(4), 395–409. https://doi.org/1.1037/a0020862

Fischer-Kern, M., Fonagy, P., Kapusta, N. D., Luyten, P., Boss, S., Naderer, A., Blüml, V., & Leithner, K. (2013). Mentalizing in female inpatients with major depressive disorder.

Journal of Nervous and Mental Disease, 201(3), 201–207. https://doi.org/1.1097/NMD.0b013e3182845c0a

Fonagy, P. (1991). Thinking about thinking: Some clinical and theoretical considerations in the treatment of a borderline patient. *The International Journal of Psychoanalysis*, *72*(4), 639–656.

Fonagy, P., & Allison, E. (2012). What is mentalization? The concept and its foundations in developmental research. In N. Midgley & I. Vrouva (Eds.), *Minding the child: Mentalization-based interventions with children, young people and their families* (pp. 11–34). Routledge/Taylor & Francis Group.

Fonagy, P., Gergely, G., Jurist, E. L., & Target, M. (2002). *Affect regulation, mentalization and the development of the self.* New York, NY: Other Press.

Fonagy, P., & Higgitt, A. (1989). A developmental perspective on borderline personality disorder. *Revue Internationale de Psychopathologie, 1,* 125-159.

Fonagy, P., Leigh, T., Steele, M., Steele, H., Kennedy, R., Mattoon, G., Target, M., & Gerber, A. (1996). The relation of attachment status, psychiatric classification, and response to psychotherapy. *Journal of Consulting and Clinical Psychology, 64*(1), 22–31. https://doi.org/1.1037/0022-006X.64.1.22

Fonagy, P., & Luyten, P. (2016). A multilevel perspective on the development of borderline personality disorder. In D. Cicchetti D. (Ed.), *Developmental psychopathology. Vol 3: Maladaptation and psychopathology* (726 - 792). New York, NY; John Wiley & Sons.

Fonagy, P., Luyten, P., & Bateman, A. (2015). Translation: Mentalizing as treatment target in borderline personality disorder. *Personality Disorders: Theory, Research, and Treatment, 6*(4), 380–392. https://doi.org/1.1037/per0000113

Fonagy, P., Luyten, P., Moulton-Perkins, A., Lee, Y.-W., Warren, F., Howard, S., Ghinai, R., Fearon, P., & Lowyck, B. (2016). Development and validation of a self-report measure of mentalizing: The Reflective Functioning Questionnaire. *PLoS ONE, 11*(7), e0158678. https://doi.org/1.1371/journal.pone.0158678

Fonagy, P., Steele, M., Steele, H., Higgitt, A., & Target, M. (1994). The Emanuel Miller Memorial Lecture 1992: The theory and practice of resilience. *Child Psychology & Psychiatry & Allied Disciplines, 35*(2), 231–257. https://doi.org/1.1111/j.1469-761.1994.tb0116.x

Fonagy, P., Steele, M., Steele, H., Moran, G. S., & Higgitt, A. C. (1991). The capacity for understanding mental states: The reflective self in parent and child and its significance for security of attachment. *Infant Mental Health Journal, 12*(3), 201–218. https://doi.org/1.1002/1097-0355(199123)12:3<201::AID-IMHJ2280120307>3..CO;2-7

Fonagy, P., & Target, M. (1996). Playing with reality, I: Theory of mind and the normal development of psychic reality. *International Journal of Psychoanalysis, 77* (2), 217–233.

Fonagy, P., & Target, M. (1997). Attachment and reflective function: Their role in self-organization. *Development and Psychopathology, 9*(4), 679-70. doi:1.1017/S0954579497001399

Fonagy, P., Target, M., Steele, H., and Steele, M. (1998). *Reflective Functioning Manual. Version 5 for Application to Adult Attachment Interviews.* [Discussion paper, University College London]. UCL Discovery.

Fossati, A., Borroni, S., Dziobek, I., Fonagy, P., & Somma, A. (2018). Thinking about assessment: Further evidence of the validity of the Movie for the Assessment of Social Cognition as a measure of mentalistic abilities. *Psychoanalytic Psychology, 35*(1), 127–141. https://doi.org/1.1037/pap0000130

Francoeur, A., Lecomte, T., Daigneault, I., Brassard, A., Lecours, V., & Hache-Labelle, C. (2020) Social cognition as mediator of romantic breakup adjustment in young adults who experienced childhood maltreatment. *Journal of Aggression, Maltreatment & Trauma*, *29* (9), 1125-1142, https://doi.org/1.1080/10926771.2019.1603177

Gagliardini, G., Gullo, S., Caverzasi, E., Boldrini, A., Blasi, S., & Colli, A. (2018). Assessing mentalization in psychotherapy: First validation of the Mentalization Imbalances Scale. *Research in Psychotherapy: Psychopathology, Process and Outcome*, *21*(3), 164–177. https://doi.org/1.4081/ripppo.2018.339

Gagnon, G. (2020). *Attachment, exploration, and internalized homonegativity* (27959038). [Doctoral dissertation, City University of New York]. ProQuest Dissertations & Theses.

George, C., Kaplan, N., & Main, M. (1984/1985/1996). *The Berkeley Adult Attachment Interview*. Unpublished manuscript, Berkeley, CA.

Gilbert, J. L. (2008). *Reflective functioning and caregiver behavior: Development of caregiver reflective functioning scales (CRFS) for use with the Circle of Security Intervention (COSI).* [Doctoral dissertation, James Madison University]. ProQuest Dissertations & Theses Global.

Gori, A., Arcioni, A., Topino, E., Craparo, G., & Lauro Grotto, R. (2021). Development of a new measure for assessing mentalizing: The Multidimensional Mentalizing Questionnaire

(MMQ). *Journal of Personalized Medicine, 11*(4), *305.* https://doi.org/1.3390/jpm11040305

Graf, E. P. (2010). *The relationship of reflective functioning and severity of agoraphobia in the outcome of a psychoanalytic psychotherapy for panic disorder*. [Doctoral dissertation, City University of New York]. ProQuest Dissertations & Theses.

Green, J., Berry, K., Danquah, A., & Pratt, D. (2021). Attachment security and suicide ideation and behavior: The mediating role of reflective functioning. *International Journal of Environmental Research and Public Health, 18*(6), 309. https://doi.org/1.3390/ijerph18063090

Grienenberger, J., Kelly, K., & Slade, A. (2005). Maternal reflective functioning, mother-infant affective communication, and infant attachment: Exploring the link between mental states and observed caregiving behavior in the intergenerational transmission of attachment. *Attachment & Human Development, 7*(3), 299–311. https://doi.org/1.1080/14616730500245963

Griva, F., Pomini, V., Gournellis, R., Doumos, G., Thomakos, P., & Vaslamatzis, G. (2020). Psychometric properties and factor structure of the Greek version of the Reflective Functioning Questionnaire. *Psychiatriki, 31*(3), 216–224. https://doi.org/1.22365/jpsych.202.313.21

Gullestad, F. S., Johansen, M. S., Høglend, P., Karterud, S., & Wilberg, T. (2013). Mentalization as a moderator of treatment effects: Findings from a randomized clinical trial for personality disorders. *Psychotherapy Research, 23*(6), 674–689. https://doi.org/1.1080/10503307.2012.684103

Gullestad, F. S., & Wilberg, T. (2011). Change in reflective functioning during psychotherapy—

A single-case study. *Psychotherapy Research, 21*(1), 97–111. https://doi.org/1.1080/10503307.201.525759

Ha, C., Sharp, C., Ensink, K., Fonagy, P., & Cirino, P. (2013). The measurement of reflective function in adolescents with and without borderline traits. *Journal of Adolescence, 36*, 1215-1223. https://doi.org/1.1016/j.adolescence.2013.09.008.

Hanak, N. (2004). Construction of a new instrument for assessment of adult and adolescent attachment—QAA. *Psychology, 37*(1), 123–142. https://doi.org/1.2298/PSI0401123H

Handeland, T. B., Kristiansen, V. R. (2017). *Certain and Uncertain Reflective Functioning in Mothers with Substance Use Disorder: Investigating the Associations between Reflective Functioning, Trauma and Executive Functions*. [Master's thesis, University of Oslo].

Handeland, T. B., Kristiansen, V. R., Lau, B., Håkansson, U., & Øie, M. G. (2019). High degree of uncertain reflective functioning in mothers with substance use disorder. *Addictive Behaviors Report, 10*, 100193. https://doi.org/1.1016/j.abrep.2019.100193

Harris, P. L. (1992). From simulation to folk psychology: The case for development. *Mind & Language, 7*(1-2), 120-144. https://doi.org/1.1111/j.1468-0017.1992.tb00201.x

Hausberg, M. C., Schulz, H., Piegler, T., Happach, C. G., Klöpper, M., Brütt, A. L., Sammet, I., & Andreas, S. (2012). Is a self-rated instrument appropriate to assess mentalization in patients with mental disorders? Development and first validation of the Mentalization Questionnaire (MZQ). *Psychotherapy Research, 22*(6), 699–709. https://doi.org/1.1080/10503307.2012.709325

Hayden, M. C., Müllauer, P. K., Gaugeler, R., Senft, B., & Andreas, S. (2018). Improvements in mentalization predict improvements in interpersonal distress in patients with mental

disorders. *Journal of Clinical Psychology, 74*(12), 2276–2286. https://doi.org/1.1002/jclp.22673

Henry, A., Allain, P., & Potard, C. (2022). Relationships between Theory of Mind and Attachment Styles in Emerging Adulthood. *Journal of Adult Development,* 1-13.

Hezel, D. M., & McNally, R. J. (2014). Theory of mind impairments in social anxiety disorder. *Behavior Therapy, 45*(4), 530–54. https://doi-org.1.1016/j.beth.2014.02.010

Horst, P. (1941). The role of predictor variables which are independent of the criterion. *Social Science Research Council Bulletin, 48*, 431-436.

Huang, J. L., Curran, P.G., Keeney, J., Poposki, E.M., & DeShon, R.P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology, 27*, 99–114. https://doi.org/1.1007/s10869-011-9231-8

Huang, Y. L., Fonagy, P., Feigenbaum, J., Montague, P. R., Nolte, T., & London Personality and Mood Disorder Research Consortium. (2020). Multidirectional pathways between attachment, mentalizing, and posttraumatic stress symptomatology in the context of childhood trauma. *Psychopathology, 53*(1), 48–58. https://doi.org/1.1159/000506406

Innamorati, M., Imperatori, C., Harnic, D., Erbuto, D., Patitucci, E., Janiri, L., Lamis, D. A., Pompili, M., Tamburello, S., & Fabbricatore, M. (2017). Emotion regulation and mentalization in people at risk for food addiction. *Behavioral Medicine, 43*(1), 21–3. https://doi.org/1.1080/08964289.2015.1036831

Jańczak, M. (2021). Polish adaptation and validation of the Mentalization Scale (MentS)—A self-report measure of mentalizing. *Psychiatria Polska, 55*(6), 1257-1274.

Johnston, L., Miles, L., & McKinlay, A. (2008). A critical review of the eyes test as a measure of social-cognitive impairment. *Australian Journal of Psychology, 60*(3), 135-141. https://doi.org/1.1080/00049530701449521

Jović, M. G., & Jovančević, A. Z. (2019). Capacity to mentalize and achievement motivation. *Proceedings of the Faculty of Teacher Education, 13,* 109–121.

Jurist, E .L. (2005). Mentalized affectivity. *Psychoanalytic Psychology, 22*(3), 426 – 444. https://doi.org/1.1037/0736-9735.22.3.426

Kaiser, H. F., & Rice, J. (1974). Little Jiffy, Mark IV. *Educational and Psychological Measurement, 34*(1), 111–117. https://doi.org/1.1177/001316447403400115

Katznelson, H. (2014). Reflective functioning: A review. *Clinical Psychology Review, 34*(2), 107–117. https://doi.org/1.1016/j.cpr.2013.12.003

Karlsson, R., & Kermott, A. (2006). Reflective- functioning during the process in brief psychotherapies. *Psychotherapy: Theory, Research, Practice, Training, 43*(1), 65–84. https://doi.org/1.1037/0033-3204.43.1.65

Kayha, Y., & Munguldar, K. (2022). Difficulties in emotion regulation mediated the relationship between reflective functioning and borderline personality symptoms among non-clinical adolescents. *Psychological Reports*, 1- 2. https://doi.org/1.1177/00332941211061072

Langner, T. F. (1962). A 22-Item screening score of psychiatric symptoms indicating impairment. *Journal of Health and Human Behavior, 3*, 269-276.

Levy, K. N., Meehan, K. B., Kelly, K. M., Reynoso, J. S., Weber, M., Clarkin, J. F., & Kernberg, O. F. (2006). Change in attachment patterns and reflective function in a randomized control trial of transference-focused psychotherapy for borderline personality disorder.

Journal of Consulting and Clinical Psychology, 74(6), 1027–104. https://doi.org/1.1037/0022-006X.74.6.1027

Li, E. T., Carracher, E., & Bird, T. (2020). Linking childhood emotional abuse and adult depressive symptoms: The role of mentalizing incapacity. *Child Abuse & Neglect, 99*, 104253. https://doi.org/1.1016/j.chiabu.2019.104253

Luyten, P., Campbell, C., & Fonagy, P. (2020). Borderline personality disorder, complex trauma, and problems with self and identity: A social-communicative approach. Journal of Personality. 88 (1), 88– 105. https://doi.org/1.1111/jopy.12483

Luyten, P., Campbell, C., Allison, E., & Fonagy, P. (2020). The mentalizing approach to psychopathology: State of the art and future directions. *Annual Review of Clinical Psychology, 16*, 297–325. https://doi.org/1.1146/annurev-clinpsy-071919-015355

Luyten, P., Fonagy, P., Lowyck, B., & Vermote, R. (2012). The assessment of mentalization. In A. Bateman & P. Fonagy (Eds.), *Handbook of mentalizing in mental health practice* (pp. 43-65). *Washington, DC: American Psychiatric Association*.

MacBeth, A., Gumley, A., Schwannauer, M., & Fisher, R. (2011). Attachment states of mind, mentalization, and their correlates in a first-episode psychosis sample. *Psychology and Psychotherapy: Theory, Research and Practice, 84*(1), 42–57. https://doi.org/1.1348/147608310X530246

McGowan, N. M., Syam, N., McKenna, D., Pearce, S., & Saunders, K. E. (2021). A service evaluation of short-term mentalization based treatment for personality disorder. *BJPsych Open, 7*, 1-8. https://doi.org/1.1192/bjo.2021.974

McLaren, V., Gallagher, M., Hopwood, C. J., & Sharp, C. (2022). Hypermentalizing and borderline personality disorder: a meta-analytic review. *American Journal of Psychotherapy, 75*(1), 21-31. https://doi.org/1.1176/appi.psychotherapy.20210018

Meehan, K. B., Levy, K. N., Reynoso, J. S., Hill, L. L., & Clarkin, J. F. (2009). Measuring reflective function with a multidimensional rating scale: Comparison with scoring reflective function on the AAI. *Journal of the American Psychoanalytic Association, 57*(1), 208–213. https://doi.org/1.1177/00030651090570011008

Mikolajczak, M., Nelis, D., Hansenne, M., & Quoidbach, J. (2008). If you can regulate sadness, you can probably regulate shame: Associations between trait emotional intelligence, emotion regulation and coping efficiency across discrete emotions. *Personality and Individual Differences, 44,* 1356–1368. doi:10.1016/j.paid.2007.12.004

Montag, C., Dziobek, I., Richter, I. S., Neuhaus, K., Lehmann, A., Sylla, R., Heekeren, H. R., Heinz, A., & Gallinat, J. (2011). Different aspects of theory of mind in paranoid schizophrenia: Evidence from a video-based assessment. *Psychiatry Research, 186*(2-3), 203–209. https://doi-org.1.1016/j.psychres.201.09.006

Morandotti, N., Brondino, N., Merelli, A., Boldrini, A., De Vidovich, G.Z., Ricciardo, S., Abbiati, V., Ambrosi, P., Caverzasi, E., Fonagy, P., & Luyten, P. (2018) The Italian version of the Reflective Functioning Questionnaire: Validity data for adults and its association with severity of borderline personality disorder. *PLoS ONE 13*(11): e0206433. https://doi.org/1.1371/journal.pone.0206433

Mousavi, P. S., Vahidi, E., Ghanbari, S., Khoshroo, S., & Sakkaki, S. Z. (2021). Reflective Functioning Questionnaire (RFQ): Psychometric properties of the Persian translation and exploration of its mediating role in the relationship between attachment to parents and internalzing and externalizing problems in adolescents. *Journal of Infant, Child & Adolescent Psychotherapy, 20*(3), 313 33. https://doi.org/1.1080/15289168.2021.1945721

Müller, C., Kaufhold, J., Overbeck, G., & Grabhorn, R. (2006). The importance of reflective functioning to the diagnosis of psychic structure. *Psychology and Psychotherapy: Theory, Research and Practice, 79*(4), 485–494. https://doi.org/1.1348/147608305x68048

Müller, S., Wendt, L. P., Spitzer, C., Masuhr, O., Back, S. N., & Zimmermann, J. (2021). A critical evaluation of the reflective functioning questionnaire (RFQ). *Journal of Personality Assessment,* 1-15. https://doi.org/1.1080/00223891.2021.1981346

Mikulincer, M., Shaver, P. R., & Pereg, D. (2003). Attachment theory and affect regulation: The dynamics, development, and cognitive consequences of attachment-related strategies. *Motivation and Emotion, 27*(2), 77–102. https://doi.org/10.1023/A:1024515519160

Murphy, B. A., & Lilienfeld, S. O. (2019). Are self-report cognitive empathy ratings valid proxies for cognitive empathy ability? Negligible meta-analytic relations with behavioral task performance. *Psychological Assessment, 31*(8), 1062–1072. https://doi.org/1.1037/pas0000732

Nettle, D., & Liddle, B. (2008). Agreeableness is related to social-cognitive, but not social-perceptual, theory of mind. *European Journal of Personality, 22*(4), 323–335. https://doi.org/1.1002/per.672

Oakley, B. F. M., Brewer, R., Bird, G., & Catmur, C. (2016). Theory of mind is not theory of emotion: A cautionary note on the Reading the Mind in the Eyes Test. *Journal of Abnormal Psychology, 125*(6), 818–823. https://doi-org.1.1037/abn0000182

O'Brien, E. (1981). *The Self Report Inventory: Constructions and validation of a multi-dimensional measure of the self concept and sources of self esteem*. [Doctoral dissertation, University of Massachusetts]. Scholar Works University of Massachusetts

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instrumentation, and Computers, 32*, 396– 402. https://doi-org.1.3758/BF03200807

Paridaens, P. (2012). *Reliability and validity of the Mentalization Questionnaire (MZQ) in forensic care*. [Master's thesis, Tilburg University].

Pedersen, S. H., Lunn, S., Katznelson, H., & Poulsen, S. (2012). Reflective functioning in 70 patients suffering from bulimia nervosa. *European Eating Disorders Review, 20*(4), 303-31. https://doi.org/1.1002/erv.2158

Preissler, S., Dziobek, I., Ritter, K., Heekeren, H.R., & Roepke, S. (2010). Social cognition in borderline personality disorder: Evidence for disturbed recognition of emotions, thoughts, and intentions of others. *Frontiers in Behavioral Neuroscience, 4*, 1-8. https://doi.org/1.3389/fnbeh.201.00182

Ponti, L., Stefanini, M. C., Gori, S., & Smorti, M. (2019). The assessment of mentalizing ability in adolescents: The Italian adaptation of the Mentalization Questionnaire (MZQ). *TPM-Testing, Psychometrics, Methodology in Applied Psychology*, *26*(1), 29–38. https://doi.org/1.4473/TPM26.1.2

Quattropani, M. C., Geraci, A., Lenzo, V., Sardella, A., & Schimmenti, A. (2022). Failures in reflective functioning, dissociative experiences, and eating disorder: A study on a sample of Italian adolescents. *Journal of Child & Adolescent Trauma*, *15*(2), 365–374. https://doi.org/1.1007/s40653-022-00450-9

Raimondi, G., Samela, T., Lester, D., Imperatori, C., Carlucci, L., Contardi, A., Balsamo, M., & Innamorati, M. (2021). Psychometric properties of the Italian Mentalization

Questionnaire: Assessing structural invariance and construct validity. *Journal of Personality Assessment.* https://doi.org/1.1080/00223891.2021.1991362

Rothschild-Yakar, L., Stein, D., Goshen, D., Shoval, G., Yacobi, A., Eger, G., Kartin, B., & Gur, E. (2019). Mentalizing self and other and affect regulation patterns in anorexia and depression. *Frontiers in Psychology, 10,* 2223. https://doi.org/1.3389/fpsyg.2019.02223

Reniers, R. L. E. P., Corcoran, R., Drake, R., Shryane, N. M., & Völlm, B. A. (2011). The QCAE: A Questionnaire of Cognitive and Affective Empathy. *Journal of Personality Assessment, 93*(1), 84–95. https://doi.org/1.1080/00223891.201.528484

Richter, F., Steinmair, D. & Löffler-Stastka, H. (2021). Construct validity of the Mentalization Scale (MentS) within a mixed psychiatric sample. *Frontiers in Psychology, 12*, 608214. https://doi.org/1.3389/fpsyg.2021.608214

Rifkin-Zybutz, R., Moran, P., Nolte, T., Feigenbaum, J., King-Casas, B., Fonagy, P., & Montague, R. (2021). Impaired mentalizing in depression and the effects of borderline personality disorder on this relationship. *Borderline Personality Disorder and Emotion Dysregulation, 8*(15), 1–6. https://doi.org/1.1186/s40479-021-00153-x

Ripoll, L. H., Snyder, R., Steele, H., & Siever, L. J. (2013). The neurobiology of empathy in borderline personality disorder. *Current Psychiatry Reports, 15* (3), 344. https://doi.org/1.1007/s11920-012-0344-1

Rishede, M.Z., Juul, S., Bo, S., Gondan, M., Bjerrum Møeller, S., & Simonsen, S. (2021). Personality functioning and mentalizing in patients with subthreshold or diagnosed borderline personality disorder: Implications for ICD-11. *Frontiers in Psychiatry, 12*, 1 – 1. https://doi.org/1.3389/fpsyt.2021.634332

Rosenblum, K. L., Mcdonough, S. C., Sameroff, A. J., & Muzik, M. (2008). Reflection in thought and action: Maternal parenting reflectivity predicts mind-minded comments and interactive behavior. *Infant Mental Health Journal, 29*(4), 362–376. https://doi.org/1.1002/imhj.20184

Rosso, A. M., Viterbori, P., & Scopesi, A. M. (2015). Are maternal reflective functioning and attachment security associated with preadolescent mentalization? *Frontiers in Psychology, 6*, 1134. https://doi.org/1.3389/fpsyg.2015.01134

Rudden, M. G., Milrod, B., Target, M., Ackerman, S., & Graf, E. (2006). Reflective functioning in panic disorder patients: A pilot study. *Journal of the American Psychoanalytic Association, 54*(4), 1339–1343. https://doi.org/1.1177/00030651060540040109

Rutherford, H. J. V., Booth, C. R., Luyten, P., Bridgett, D. J., & Mayes, L. C. (2015). Investigating the association between parental reflective functioning and distress tolerance in motherhood. *Infant Behavior & Development, 40*, 54–63. https://doi.org/1.1016/j.infbeh.2015.04.005

Rutimann, D. D., & Meehan, K. B. (2012). Validity of a brief interview for assessing reflective function. *Journal of the American Psychoanalytic Association, 60*(3), 577–589. https://doi.org/1.1177/0003065112445616

Sebastian, C. L., Fontaine, N. M. G., Bird, G., Blakemore, S.-J., De Brito, S. A., McCrory, E. J. P., & Viding, E. (2012). Neural processing associated with cognitive and affective Theory of Mind in adolescents and adults. *Social Cognitive and Affective Neuroscience, 7*(1), 53–63. https://doi-org/1.1093/scan/nsr023

Sacchetti, S., Robinson, P., Bogaardt, A., Clare, A., Ouellet-Courtois, C., Luyten, P., Bateman, A., & Fonagy, P. (2019). Reducing mentalizing in patients with bulimia

nervosa and features of borderline personality disorder: A case-control study. *BMC Psychiatry, 19,* 134. https://doi-org/1.1186/s12888-019-2112-9

Safran, J. D., & Muran, J. C. (2007). *Therapist relationship interview.* Unpublished manuscript.

Safran, J. D., Muran, J. C., & Shaker, A. (2014). Research on therapeutic impasses and ruptures in the therapeutic alliance. *Contemporary Psychoanalysis, 50*(1-2), 211–232. https://doi.org/1.1080/0010753.2014.880318

Schechter, D. S., Coates, S. W., Kaminer, T., Coots, T., Zeanah, C. H., Jr., Davies, M., Schonfeld, I. S., Marshall, R. D., Liebowitz, M. R., Trabka, K. A., McCaw, J. E., & Myers, M. M. (2008). Distorted maternal mental representations and atypical behavior in a clinical sample of violence-exposed mothers and their toddlers. *Journal of Trauma & Dissociation, 9*(2), 123–147. https://doi.org/1.1080/15299730802045666

Schwarzer, N.-H., Nolte, T., Fonagy, P., & Gingelmaier, S. (2021a). Mentalizing and emotion regulation: Evidence from a nonclinical sample. *International Forum of Psychoanalysis, 30*(1), 34–45. https://doi.org/1.1080/0803706X.2021.1873418

Schwarzer, N.-H., Nolte, T., Fonagy, P., & Gingelmaier, S. (2021b). Mentalizing mediates the association between emotional abuse in childhood and potential for aggression in non -clinical adults. *Child Abuse & Neglect, 115,* 105018. https://doi.org/1.1016/j.chiabu.2021.105018

Shamay-Tsoory, S. G., Harari, H., Aharon-Peretz, J., & Levkovitz, Y. (2010). The role of the orbitofrontal cortex in affective theory of mind deficits in criminal offenders with psychopathic tendencies. *Cortex: A Journal Devoted to the Study of the Nervous System and Behavior, 46*(5), 668–677. https://doi-org/1.1016/j.cortex.2009.04.008

Sharp, C., Pane, H., Ha, C., Venta, A., Patel, A. B., Sturek, J., & Fonagy, P. (2011). Theory of mind and emotion regulation difficulties in adolescents with borderline traits. *Journal of the American Academy of Child & Adolescent Psychiatry, 50*(6), 563–573. https://doi.org/1.1016/j.jaac.2011.01.017

Sharp, C., Steinberg, L., McLaren, V., Weir, S., Ha, C., & Fonagy, P. (2021). Refinement of the Reflective Function Questionnaire for Youth (RFQY) Scale B using item response theory. *Assessment*, 10731911211003971. Advance online publication. https://doi.org/1.1177/10731911211003971

Sharp, C., Williams, L. L., Ha, C., Baumgardner, J., Michonski, J., Seals, R., Patel, A. B., Bleiberg, E., & Fonagy, P. (2009). The development of a mentalization-based outcomes and research protocol for an adolescent inpatient unit. *Bulletin of the Menninger Clinic, 73*(4), 311–338. https://doi.org/1.1521/bumc.2009.73.4.311

Sharp, C., Pane, H., Ha, C., Venta, A., Patel, A. B., Sturek, J., & Fonagy, P. (2011). Theory of mind and emotion regulation difficulties in adolescents with borderline traits. *Journal of the American Academy of Child & Adolescent Psychiatry, 50*(6), 563–573. https://doi.org/1.1016/j.jaac.2011.01.017

Shmueli-Goetz, Y., Target, M., Fonagy, P., & Datta, A. (2008). The Child Attachment Interview: A psychometric study of reliability and discriminant validity. *Developmental Psychology, 44*, 939– 956. https://doi.org/1.1037/0012-1649.44.4.939

Skårderud, F. (2007). Eating one's words: Part 3. Mentalisation-based psychotherapy for anorexia nervosa: An outline for a treatment and training manual. *European Eating Disorders Review, 15*(5), 323–339. https://doi.org/1.1002/erv.817

Slade, A. (2003). *The Pregnancy Interview – Revised*. Unpublished manuscript. Yale Child Study Center, New Haven, CT.

Slade, A., Bernbach, E., Grienenberger, J., Levy, D. W., & Locker, A. (2004). *The Parent Development Interview and the Pregnancy Interview: Manuals for Scoring*. New Haven, CT: City College of New York and Yale Child Study Center.

Slade, A., Grienenberger, J., Bernbach, E., Levy, D., & Locker, A. (2005). Maternal reflective functioning, attachment, and the transmission gap: A preliminary study. *Attachment & Human Development, 7*(3), 283–298. https://doi.org/1.1080/14616730500245880

Slade, A., & Patterson, M. (2005). *Addendum to Reflective Functioning Scoring Manual for use with the Pregnancy Interview.* Unpublished manuscript. The Psychological Center at the City College of New York.

Sleed, M., Slade, A., & Fonagy, P. (2020). Reflective Functioning on the Parent Development Interview: Validity and reliability in relation to socio-demographic factors. *Attachment & Human Development, 22*(3), 310–331.

Solbakken, O. A., Hansen, R. S., & Monsen, J. T. (2011). Affect integration and reflective function: Clarification of central conceptual issues. *Psychotherapy Research, 21*(4), 482–496. https://doi.org/1.1080/10503307.2011.583696

Song, H., & Choi, H.A. (2017). Exploration of the Factor Structure of the Mentalization Questionnaire (MZQ) in 16–17-year-old Korean Adolescents. *Korean Journal of Clinical Psychology, 36 (3), 391- 401.* https://doi.org/1.15842/kjcp.2017.36.3.009

Spitzer, C., Zimmermann, J., Brähler, E., Euler, S., Wendt, L. P., & Müller, S. (2020). The German version of the Reflective Functioning Questionnaire (RFQ): German general

population. *Psychotherapie -Psychosomatik-Medizinische Psychologie, 71*(3-04), 124–131. https://doi.org/1.1055/a-1234-6317

Stacks, A. M., Muzik, M., Wong, K., Beeghly, M., Huth-Bocks, A., Irwin, J. L., & Rosenblum, K. L. (2014). Maternal reflective functioning among mothers with childhood maltreatment histories: Links to sensitive parenting and infant attachment security. *Attachment & Human Development, 16*(5), 515–533. https://doi.org/1.1080/14616734.2014.935452

Stagaki, M., Nolte, T., Feigenbaum, J., King-Casas, B., Lohrenz, T., Fonagy, P., Personality and Mood Disorder Research Consortium, & Montague, P. R. (2022). The mediating role of attachment and mentalising in the relationship between childhood maltreatment, self-harm and suicidality. *Child Abuse & Neglect, 128*, 105576. https://doi.org/1.1016/j.chiabu.2022.105576

Stanojević, T.S., Radev, M.T., & Bogdanović, A. (2020). From preoccupied attachment to depression: Serial mediation model effects on a sample of women. *Annual of Social Work, 27* (3), 523–542. https://doi.org/1.3935/ljsr.v27i1.334

Staun, L., Kessler, H., Buchheim, A., Kächele, H., & Taubner, S. (2010). Mentalization and chronic depression. *Psychotherapeut, 55*(4), 299–305.

Taubner, S., Hörz, S., Fischer-Kern, M., Doering, S., Buchheim, A., & Zimmermann, J. (2013a). Internal structure of the Reflective Functioning Scale. *Psychological Assessment, 25*(1), 127–135. https://doi.org/1.1037/a0029138

Taubner, S., Kessler, H., Buchheim, A., Kächele, H., & Staun, L. (2011). The role of mentalization in the psychoanalytic treatment of chronic depression. *Psychiatry:*

Interpersonal and Biological Processes, *74*(1), 49–57. https://doi.org /1.1521/psyc.2011.74.1.49

Taubner, S., White, L. O., Zimmermann, J., Fonagy, P., & Nolte, T. (2013b). Attachment-related mentalization moderates the relationship between psychopathic traits and proactive aggression in adolescence. *Journal of Abnormal Child Psychology, 41*(6), 929-938. https://doi.org/1.1007/s10802-013-9736-x

Taubner, S., Zimmermann, L., Ramberg, A., & Schröder, P. (2016). Mentalization mediates the relationship between early maltreatment and potential for violence in adolescence. *Psychopathology, 49*(4), 236 – 246. https://doi.org/1.1159/000448053

Vahidi, E., Ghanbari, S., & Behzadpoor, S. (2021). The relationship between mentalization and borderline personality features in adolescents: Mediating role of emotion regulation. *International Journal of Adolescence and Youth, 26*(1), 284–293. https://doi.org/1.1080/ 02673843.2021.1931376

Vijayaraghavan, A., Bhola, P., Thirthalli, J., & Mehta, U. M. (2018). Pattern of social cognition deficits in individuals with borderline personality disorder. *Asian Journal of Psychiatry, 33*, 105–112. https://doi.org/1.1016/j.ajp.2018.03.010

Waal, F. D. (2006). Morally evolved: Primate social instincts, human morality, and the rise and fall of "veneer theory" In F. de Waal & S. Macedo, J. Ober (Eds.), *Primates and philosophers: How morality evolved* (pp. 4–58). Princeton University Press.

Wallin, D. J. (2007). Mary Main: Mental Representations, Metacognition, and the Adult Attachment Interview. In *Attachment in Psychotherapy* (pp.40-41). New York: Guilford Press.

Ward, A., Ramsay, R., Turnbull, S., Steele, M., Steele, H., & Treasure, J. (2001). Attachment in anorexia nervosa: A transgenerational perspective. *British Journal of Medical Psychology, 74*(4), 497–505. https://doi.org/1.1348/000711201161145

Woynowskie, K. (2015). *Mentalization and overlapping constructs: Mindfulness, empathy, emotional intelligence, psychological mindedness using exploratory factor analysis*. [Doctoral dissertation, The Chicago School of Professional Psychology]. ProQuest Dissertations & Theses.

Woźniak-Prus, M., Gambin, M.., Cudo, A., & Sharp, C. (2022): Investigation of the Factor Structure of the Reflective Functioning Questionnaire (RFQ-8): One or Two Dimensions? *Journal of Personality Assessment*, 1–11. Advance online publication. https://doi.org/1.1080/00223891.2021.2014505

Wimsatt, W. C. (1986). Developmental constraints, generative entrenchment, and the innate-acquired distinction. In W. Bechtel (Ed.), *Integrating scientific disciplines. Science and philosophy* (pp. 185–208). Springer.

Winkler, A. (2014). Resilience as reflexivity: A new understanding for work with looked-after children. *Journal of Social Work Practice, 28*(4), 461–478. https://doi.org/ 1.1080/02650533.2014.896784

Wu, H., Fung, B. J., & Mobbs, D. (2022). Mentalizing during social interaction: The development and validation of the Interactive Mentalizing Questionnaire. *Frontiers in Psychology, 12*, 791835. https://doi.org/1.3389/fpsyg.2021.791835

Zalla, T., Miele, D., Leboyer, M., & Metcalfe, J. (2015). Metacognition of agency and theory of mind in adults with high functioning autism. *Consciousness and Cognition: An International Journal, 31*, 126–138. https://doi.org/1.1016/j.concog.2014.11.001

Zeanah, C. H., & Benoit, D. (1995). Clinical applications of a parent perception interview in infant mental health. *Child and Adolescent Clinics of North America, 4*(3), 539–554. https://doi.org/1.1016/S1056-4993(18)30418-8